AD_____

Grant Number DAMD17-94-J-4354

TITLE: Compression and Classification of Digital Mammograms for Storage, Transmission, and Computer Aided Screening

PRINCIPAL INVESTIGATOR: Robert M. Gray, Ph.D.

CONTRACTING ORGANIZATION: Stanford University
Stanford, California 94305

REPORT DATE: March 1997

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 1997	Final (1 Aug 94 - 31 Jan 97)

4. TITLE AND SUBTITLE
Compression and Classification of Digital Mammograms for Storage, Transmission, and Computer Aided Screening

5. FUNDING NUMBERS
DAMD17-94-J-4354

6. AUTHOR(S)
Robert M. Gray, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Stanford University
Stanford, California 94305-9510

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200*

The substitution of digital representations for standard film/screen mammography analog images provides access to methods for digital storage and transmission and enables the use of a variety of digital image processing techniques, including enhancement and computer assisted detection and classification of abnormalities. Lossy image compression can further improve the efficiency of transmission and storage and can facilitate subsequent image processing. Digitization, digital acquisition, lossy compression, and enhancement and highlighting schemes all alter an image from its traditional form and it is critically important that any such alteration be clearly demonstrated to improve or at least not damage the utility of the image. The primary goal of this project was the development of a clinical experimental protocol for evaluating the comparative quality of analog, digital, and lossy compressed digital mammograms in a manner not suffering from the fundamental flaws of traditional receiver operating characteristic (ROC) methods. Such a protocol was developed and implemented in a pilot experiment involving 57 patients and six radiologist judges interpreting original, 12 bits per pixel (bpp) digitized, and lossy compressed digital mammograms down to .15 bpp. The protocol, experiment, and results are described along with related algorithms for compression and combined compression and classification.

14. SUBJECT TERMS Breast Cancer , Digital Mammography, Lossy Data Compression, Image Quality Evlauation, Automated Classification

15. NUMBER OF PAGES 64

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

X For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

Robert M. Gray 27 February 1997

PI - Signature Date

Contents

1 Introduction

1.1 Background

X-ray-mammography is the most sensitive widely used technique for detecting breast cancer [1], with a reported sensitivity of 85–95% for detecting small lesions. Most non-invasive ductal carcinomas, or DCIS, are characterized by tiny non-palpable calcifications detected at screening mammography [2, 3, 4]. Traditional mammography is essentially analog photography using X-rays in place of light and analog film for display. The digital format is required for access to modern digital storage, transmission, and digital computer processing. Images in analog format are not easily distributed to multiple sites, either in-hospital or off-site; and there is the cost of personnel salary and benefits to store, archive, and retrieve the films. Currently only 30% of American women get regular mammograms, and the storage problems will be compounded if this number increases with better education or wider insurance coverage. Digital image processing provides the possibilities for easy image retrieval, efficient storage, rapid image transmission for off-site diagnoses, and the maintenance of large image banks for purposes of teaching and research. It allows filtering, enhancement, classification, and combining images obtained from different modalities, all of which can assist screening, diagnosis, research, and treatment. Retrospective studies of interval cancers (carcinomas detected in the time intervals between mammographic screenings which were interpreted as normal) show that observer error can comprise up to 10% of such cancers. That is to say, carcinomas present on the screening mammograms were missed by the radiologist because of fatigue, misinterpretation, distraction, obscuration by a dense breast, or other reasons [5, 6, 7]. To this end, schemes for computer-aided diagnosis (CAD) may assist the radiologist in the detection of clustered microcalcifications and masses [8, 9, 10, 11, 12]. Virtually all existing CAD schemes require images in digital format.

To take advantage of digital technologies, either analog signals such as X-rays must be converted into a digital format, or they must be directly acquired in digital form. Digitization of an analog signal causes a loss of information and a possible deterioration of the signal. In addition, with the increasing accuracy and resolution of analog-to-digital converters, the quantities of digital information produced can overwhelm available resources. A typical digitized mammogram with 4500×3200 picture elements (pixels) with 50 micron spot size and 12 bit per pixel depth requires approximately 38 megabytes of data. Complete studies can easily require unacceptably long transmission times through crowded digital networks and can cause serious data management problems in local disk storage. Advances in technologies for transmission and storage do not alone solve the problem. Compression is desirable and often essential for efficiency of cost and time for storage and communication.

A digital compression system typically consists of a signal decomposition such as Fourier or wavelet, a quantization operation on the coefficients, and finally lossless or entropy coding such as Huffman or arithmetic coding. Decompression reverses the above process; although if quantization is used, the system will be lossy in the sense that the image will not be perfectly reconstructible from the digital representation. Quantization is only approximately reversible. Lossless or invertible coding allows perfect reconstruction of a digital image, but typically yields compression ratios of only 2:1 to 3:1 on still frame grayscale medical images. This modest compression is often inadequate. Lossy coding can provide excellent quality at a fraction of the bit rate [13, 14, 15, 16, 17]. The *bit rate* of a compression system is the average number of bits produced by the encoder for each image pixel. If the original image has 12 bits per pixel (bpp) and the compression algorithm has rate R bpp, then the *compression ratio* is $12{:}R$.

Early studies of lossy compressed medical images performed compression using variations on the standard discrete cosine transform (DCT) coding algorithm combined with scalar quantization and lossless coding. These are variations of the international standard Joint Photographic Experts Group (JPEG) compression algorithm [18, 19, 20]. The American College of Radiology–National Electrical Manufacturers Association (ACR-NEMA) standard [21] has not yet firmly recommended a specific compression scheme, but transform coding methods are suggested.

More recent studies of efficient lossy image compression algorithms have used subband or wavelet decompositions combined with scalar or vector quantization [22, 23, 24, 25, 26, 27, 28, 29]. These signal decompositions provide several potential advantages over traditional Fourier-type decompositions, including better concentration of energy, better decorrelation for a wider class of signals, better basis functions for images than the smoothly oscillating sinusoids of Fourier analysis because of diminished Gibbs and edge effects, and better localization in both time and frequency. Because of their sliding-block operation using 2-dimensional linear filters, they do not produce blocking artifacts.

Analog mammography remains the gold standard against which all other imaging modalities must be judged. In a medical application it does not suffice for an image to simply "look good" or to have a high signal-to-noise ratio (SNR), nor should one necessarily require that original and processed images be visually indistinguishable. Rather it must be convincingly demonstrated that essential information has not been lost and that the processed image is at least of equal utility for interpretation as the original. Image quality is typically quantified objectively by average distortion or SNR, and subjectively by statistical analyses of viewers' scores on quality (e.g., analysis of variance (ANOVA) and receiver operating characteristic (ROC) curves). Examples of such approaches may be found in [30, 15, 31, 32, 14, 13, 33].

ROC analysis is the dominant technique for evaluating the suitability of radiologic techniques for real applications [34, 35, 36, 37]. Its origins are in the theory of signal detection: a filtered version of signal plus Gaussian noise is sampled and compared to a threshold. If the threshold is exceeded, then the signal is said to be there. As the threshold varies, the probability of erroneously declaring a signal absent and the probability of erroneously declaring a signal there when it is not also vary, and in opposite directions. The plotted curve is a summary of the tradeoff in these two quantities; more precisely, it is a plot of *true positive rate* or *sensitivity* against *false positive rate*, the complement of *specificity*. Summary statistics, such as the area under the curve, can be used to summarize overall quality in terms of detection accuracy. In typical implementations, radiologists or other users are asked to assign integer confidence ratings to their diagnoses, and thresholds in these ratings are used in computing the curves. We have argued in our previously cited references (summarized later) that traditional ROC analysis violates several reasonable guidelines for designing experiments to measure quality and utility in medical images because of the use of artificial confidence ratings as thresholds in a binary detection problem and because of the statistical assumptions of Gaussian or Poisson behavior. In addition, traditional ROC analysis is not well suited to the study of the accuracy of detection and location when a variety of abnormalities are possible. Although extensions of ROC designed to handle location and multiple lesions have been proposed [38, 39], they inherit the more fundamental problems of the approach and are not widely used, although they are widely quoted in defense of ROC techniques being applied to non-binary detection problems. Traditional ROC analysis also does not come equipped to distinguish among the various possible notions of "ground truth" or "gold standard" in clinical experiments. As will be discussed further, the application of ROC analysis to the measurement of radiological image quality is fundamentally flawed and does not in fact demonstrate image quality or lack thereof in the real-world applications of diagnosis and screening. The problems can all be diminished significantly by alternative experimental design and

statistical analysis techniques, but the traditional methods continue to dominate practice.

During the past decade our group at Stanford University has worked to develop alternative approaches to evaluating the diagnostic accuracy of lossy compressed medical images (or any digitally processed medical images) that mimic ordinary clinical practice as closely as is reasonably possible and do not require special training or artificial subjective evaluations. These approaches apply naturally to the detection of multiple abnormalities and to measurement tasks, and require no assumptions of Gaussian behavior of crucial data. While some departures from ordinary practice are necessary and some additional information may be gathered because it is of potential interest, the essential goal remains the imitation of ordinary practice and the drawing of diagnostic conclusions based only on diagnostic simulations. The methods are developed in detail for CT and MR images [40, 41, 42, 43, 44]. Extensions to digital mammography were accomplished under this project with support from the USAMRMC and from Kodak, Inc. Early results are described in [45, 46, 47] and in a special issue of *Signal Processing* devoted to medical image compression [48] (preprints of which can be found at the World Wide Web site [49]). This report collects these preliminary results along with more recent results and relevant continuing research since the end of the Army grant.

Although the Army project has ended, statistical analysis of the data base generated by the grant continues with the support of a gift from Kodak, Inc. Additional results based on this analysis will be submitted later as an addendum to this report and will be available at the USAMRMC workshop to be held in fall of 1997. They will also form part of the manuscripts to be submitted for publication based on the final results.

The principal publications derivative from this project, this report, and subsequent addenda and manuscripts are or will be available in Adobe portable document format (pdf) at the project web site, `http://www-isl.stanford.edu/~gray/army.html`.

1.2 Quality Evaluation and the FDA

Radiology is increasingly digital and capable of using digital communication links, storage facilities, and image processing. Digitization of analog images and most image processing algorithms change images and may reduce their utility. A traditional approach to establish quality and utility in specific applications is to simulate the application in a carefully designed experiment, gather necessary data in a way that interferes with the simulation as little as possible, and analyze the resulting data to assess the validity of a specific hypothesis, such as "image type A is not different from image type B" in a specific diagnostic application. In the United States, new devices such as FFDM systems must receive approval from the Center for Devices and Radiological Health of the Federal Drug Administration (CDRH/FDA) if they are to be marketed commercially. In September, 1995 the CDRH/FDA called for comments on a *Draft Guidance* describing a proposed protocol for clinical studies that might lead to such approval. The protocol was designed to test the hypothesis that FFDM is at least as effective as traditional film/screen (F/S) mammography in screening applications. The original proposal required that more than 11,000 patients be studied, a requirement which we argued before the CDRH/FDA [50] was vastly excessive for adequately testing sensitivity and specificity of decisions regarding patient management. Because of the close connection of the proposed protocol and the protocols developed under this grant, our group became involved in the FDA open meetings and formally submitted comments on the proposed guidelines. Our arguments were based on the experimental and statistical methods developed under our current USAMRMC project. If adopted, the CDRH/FDA proposal would have placed an extreme burden in time and money on companies seeking approval for such devices. Such a burden could delay the application of promising new technologies, could have a chilling effect on research, and could discourage com-

panies from remaining in or entering the field. The controversy focused on fundamental issues of how to quantify quality and utility of medical images that have been altered by digitization, digital acquisition, or image processing. We believe that our testimony based on the protocol and analysis methods that we developed had an impact on the subsequent CDRH/FDA *Final Guidance* of June 1996 that removed the 11,000 patient requirement and proposed a study much closer to that proposed by us. Problems remained, however, as the *Guidance* also required only 2 radiologist judges in clinical experiments, in our opinion erring too far in the direction of insufficient requirements for the public's welfare. We believe there should be at least nine and possibly twelve judges in the experiment to provide adequate statistical power for the size and tasks considered. We strongly believe that research of the type performed in this project and described in this report is required in order to obtain accurate estimates of the size and power for demonstrating equivalence or superiority of management and detection tasks in mammographic image interpretation as a function of number of studies and judges, and that such information will be essential to a correct validation of the quality of images produced by digital mammography devices and of any images which are modified by computer processing, including computer assisted diagnosis and enhancement. We hope that the research described in this report will contribute to progress in image quality evaluation and will cause some researchers in the field to take into consideration the concerns raised here and the papers cited.

1.3 Compression and Classification

The study reported here used a particular compression algorithm, one which was considered by us to be the best performing available algorithm in terms of bit rate, quality, and implementational complexity. During the course of the research, however, other compression methods were investigated and systems combining compression and statistical classification were developed and evaluated by simulation. The algorithm work was jointly supported by this Grant and by National Science Foundation Grant No. NSF MIP-9016974. Although some of the compression algorithms studied were competitive with the embedded wavelet schemes in terms of bit rate/quality tradeoffs, the scheme adopted was still superior in terms of simplicity of implementation. The algorithms considered for combining compression and classification are aimed at the long term goal of simultaneously compressing and segmenting an image into suspicious and nonsuspicious regions of various types. These algorithms combine aspects of compression with empirical Bayes classification and hold promise for future applications to computer assisted screening and diagnosis, but in their current state they are not competitive in performance with most of the schemes that have been published to date. In their favor they are much simpler to implement and incorporate the necessary signal processing into the compression algorithm. Applications of this work to digital mammography were studied during this project and the results are reported briefly here.

2 Hypothesis/Purpose

The nominal hypothesis of this project has been that digitized mammograms and lossy compressed digitized mammograms are at least as good as traditional film/screen mammography for the detection of abnormalities, provided that the bit rate is sufficient. We believe that .25 bpp (in place of 12 bpp originals at 50 micron spot sizes) are sufficient for this purpose. The more fundamental hypothesis is that the nominal hypothesis can be tested by using a protocol consistent with basic principles of good experimental design. The clinical experiments fulfill the underlying goals of the proposed CDRH/FDA protocol for digital vs. analog comparisons.

The pilot study and the results reported here are consistent with these hypotheses, but the study is too small to be definitive. The issue of the number of patients and radiologists needed to provide the size and power for the statistical tests to be definitive is considered in this report. Nonetheless, the experiment reported here was in our knowledge the largest experiment to date concerning the comparison of analog, digital, and lossy compressed digital radiological images in terms of the quantity of data gathered.

The protocol developed here is applicable to comparing any two distinct image modalities used for a common purpose and hence can also be applied to tasks such as quantifying the effects of softcopy interpretation and computer assisted diagnostic methods involving image enhancement, image segmentation, and computer inserted clues.

The conservative hypothesis of substantial equality used here was chosen as that is the criterion demanded by the FDA for the validation of digital mammography devices for screening applications — they must perform at least as well as traditional analog mammography. It was our view that such devices must first be shown to be no worse than existing image modalities in order to encourage further research and development of future applications which, we believe, will ultimately prove digital mammography to be in fact significantly superior.

3 Study Design

3.1 Principles of Experimental Design

The general methods used are extensions to digital mammography and elaborations of techniques developed for CT and MR images by our group and reported in [40, 41, 42, 43, 44], where all details regarding the data, compression code design, clinical simulation protocols, and statistical analyses may be found. We here describe the extensions of these methods to digital mammography. (See also [45, 46, 47].)

The general goals and specific implementation of the experimental design were developed through active cooperation among radiologists, statisticians, and electrical engineers. In addition to the Investigators, many participating volunteer radiologists (two at Stanford University, three at the University of Virginia and one at the University of California at San Francisco) contributed to the development through feedback before and during the judging sessions and the sessions establishing the independent "gold standards" described later. The cases for the pilot study were chosen to meet statistical criteria outlined later in the statistical analysis section. All radiologists were Board certified practicing mammographers.

The following general principles for protocol design have evolved from our earlier work. Although they may appear self-evident in hindsight, they provide a useful context for evaluating protocols for judging image quality in medical imaging applications and they represent an accumulation of over eight years of discussion and experience among electrical engineers, statisticians, radiologists, and medical physicists.

- The protocol should *simulate ordinary clinical practice as closely as possible.* In particular, participating radiologists (judges, observers) should perform in a manner that *mimics their ordinary practice* as closely as reasonably possible given the constraints of good experimental design.

- The studies should require *little or no special training* of their clinical participants.

- The clinical studies should *include examples of images containing the full range of possible findings*, all but extremely rare conditions.
- The findings should be *reportable using a subset of the American College of Radiology (ACR) Standardized Lexicon.*
- *Statistical analyses of the trial outcomes should be based on assumptions as to the outcomes and sources of error that are faithful to the clinical scenario and tasks.*
- *"Gold standards"* for evaluation of equivalence or superiority of algorithms *must be clearly defined and consistent with experimental hypotheses.*
- Careful *experimental design should eliminate or minimize any sources of bias in the data* that are due to differences between the experimental situation and ordinary clinical practice, e.g., learning effects that might accrue if a similar image is seen using separate imaging modalities.
- The number of patients should be sufficient to ensure *satisfactory size and power* for the principal statistical tests of interest.

The ROC assumptions and approach generally differ from clinical practice. Digitization of an analog image and lossy compression are not equivalent to the addition of signal-independent noise. Radiologists are not threshold detectors. Using properly designed ROC analysis to compare computer aided diagnosis (CAD) schemes is appropriate because such schemes almost always depend on a threshold, albeit in a possibly complicated way. No hard evidence exists, however, to support the contention that human radiologists behave in this way and, even if they did, that the ROC method of asking them for confidence ratings to interpret as thresholds in fact measures whatever internal threshold they might have. This limits the value of using ROC curves to draw conclusions about quality comparisons among radiologists or among images read by radiologists. Because of the need for confidence ratings, the traditional ROC approach requires special training to familiarize a radiologist with the rating system. On the statistical side, image data are not well modeled as known signals in Gaussian noise, and hence methods that rely on Gaussian assumptions are suspect. This is particularly true when Gaussian approximations are invoked to compute statistical size and power on a data set clearly too small to justify such approximations. Modern computer-intensive statistical sample reuse or simulation techniques can help get around the failures of Gaussian assumptions, but this does not address the more fundamental issues.

Traditional ROC methods are not suitable for detecting multiple lesions and their locations. Extensions of ROC such as LROC and FROC to consider location and multiple lesions have been proposed [38, 39], but the methods are cumbersome and inherit the remaining faults of ROC such as confidence ratings and Gaussian or Poisson assumptions on the data. In our view they attempt to fit the method (ROC analysis) to clinical practice in an artificial way, rather than trying to develop more natural methods for measuring how well radiologists perform ordinary clinical functions on competing image modalities. These methods have not been commonly adopted and appear to be primarily used as citations in defense of the criticism of ROC as being appropriate for only binary detection problems.

Traditional ROC analysis has no natural extension to problems of estimation or regression instead of detection. For example, measurement plays an important role in some diagnostic applications and there is no ROC analysis for measurement error.

Lastly, traditional ROC applications have often been lax in clarifying the "gold standard" used to determine when decisions are "correct," when in fact a variety of gold standards are possible,

each with its own uses and shortcomings. We focus on three definitions of diagnostic truth as a basis of comparison for the diagnoses on all lossy reproductions of that image:

Personal Each judge's readings on an original analog image are used as the gold standard for the readings of that same judge on the digitized version of that same image,

Independent formed by the agreement of the members of an independent expert panel, and

Separate produced by the results of further imaging studies (including ultrasound, spot and magnification mammogram studies), surgical biopsy, and autopsy.

The first two gold standards are usually established using the analog original films. As a result, they are extremely biased in favor of the established modality, i.e., the original analog film. Thus statistical analysis arguing that a new modality is equal to or better than the established modality will be conservative since the original modality is used to establish "ground truth." The personal gold standard is in fact hopelessly biased in favor of the analog films. It is impossible for the personal gold standard to be used to show that digital images are *better* than analog ones. If there is any component of noise in the diagnostic decision, the digital images cannot even be found equal to analog. The personal gold standard is often useful, however, for giving some indication of the diagnostic consistency of an individual judge. The independent gold standard is also biased in favor of the analog images, but not hopelessly so, as it is at least possible for the readings of an individual judge on either the digital or analog images to differ from the analog gold standard provided by the independent panel. If the independent panel cannot agree on a film, the film could be removed from the study; but this would forfeit potentially valuable information regarding difficult images. By suitable gathering of data, one can instead define several possible independent gold standards and report the statistics with respect to each. In particular, a cautious gold standard declares a finding if any of the panel do so. An alternative is that the panel designates a chair to make a final decision when there is disagreement.

Whenever a believable separate gold standard is available, it provides a more fair gold standard against which both old (analog) and new (digital, compressed digital) images can be compared.

3.2 Image Database

The image database for this USAMRMC project was generated in the Department of Radiology of the University of Virginia School of Medicine and is summarized in Table 1. The studies were digitized using a Lumisys Lumiscan 150 at 12 bpp with a spot size of 50 microns. Good quality directly acquired digital mammograms were not yet available when the experiment was begun, so digitized mammograms were used. The 57 studies included a variety of normal images and images containing benign and malignant objects.

For a definitive study these numbers would be scaled up to provide sufficient size and power for the statistical tests proposed. We proposed the specific numbers of Table 2 to the FDA [50], but the power and size considerations considered later suggest that this may be somewhat small.

We do not wish to simulate the proportion of normal images to ones containing pathology that would actually be found in a screening situation as there would be too few cases of pathology (6–8 cancers/1000 asymptomatic women screened). This issue is dealt with in the Statistical Analysis discussion.

3.3 Experimental Protocol

Images were viewed on hardcopy film on an alternator by judges in a manner that simulated ordinary screening and diagnostic practice as closely as possible, although patient histories and other image modalities were not provided. Two views were provided of each breast (CC and MLO). Each of the judges viewed all the images in an appropriately randomized order over the course of eighteen sessions. A clear overlay was provided for the judge to mark lesion and nipple location on the image without leaving a visible trace. For each image, the judge either indicated that the image was normal, or, if something was detected, had an assistant fill out the Observer Form found in Appendix B using the American College of Radiology (ACR) Standardized Lexicon by circling the appropriate answers or filling in blanks as directed. The instructions for assistants are also in the Addenda. Current versions of these forms along with a CGI web data entry form may be found at the project Web site [49]. The judges used a grease pencil to circle the detected item. The instructions to the judges specified that ellipses drawn around clusters should include all microcalcifications seen, as if making a recommendation for surgery, and outlines drawn around masses should include the main tumor as if grading for clinical staging, without including the spicules (if any) that extend outward from the mass. This corresponds to what is done in clinical practice except for the requirement that the markings be made on copies. The judges were allowed to use a magnifying glass to examine the films.

Although the judging form is not standard, the ACR Lexicon is used to report findings, and hence the judging requires no special training. The reported findings permit subsequent analysis of the quality of an image in the context of its true use, finding and describing anomalies and using them to assess and manage patients.

To confirm that each radiologist identifies and judges a specific finding, the location of each lesion is confirmed both on the clear overlay and the judging form. Many of these lesions were judged as 'A' (assessment incomplete), since it is often the practice of radiologists to obtain additional views in two distinct scenarios: (1) to confirm or exclude the presence of a finding, that is, a finding that may or may not represent a true lesion, or (2) to further characterize a true lesion, that is, to say a lesion clearly exists but is incompletely evaluated.

The judging form allows for two meanings of the 'A' code. If the judge believes that the finding is a possible lesion, this is indicated by answering "yes" to the question "are you uncertain if the finding exists?" Otherwise, if the lesion is definite, the judges should give their best management decision based on the standard two-view mammogram.

The initial question requesting a subjective rating of diagnostic utility on a scale of 1-5 is intended for a separate evaluation of the general subjective opinion of the radiologists of the images. The degree of suspicion registered in the Management portion also provides a subjective rating, but this one is geared towards the strength of the opinion of the reader regarding the cause of the management decision. It is desirable that obviously malignant lesions in a gold standard should also be obviously malignant in the alternative method.

3.4 Statistical Analysis

Management

We first focus on patient management, the decisions that are made based on the radiologists' reading of the image. Lesion-by-lesion accuracy of detection is considered later.

Management is a key issue in digital mammography. There is concern that artifacts could be introduced, leading to an increase in false positives and hence in unnecessary biopsies. The

management categories we emphasize are the following four, given in order of increasing seriousness:

RTS incidental, negative, or benign with return to screening

F/U probably benign but requiring six month follow-up

C/B call back for more information, additional assessment needed

BX Immediate biopsy.

These categories are formed by combining categories from the basic form of Appendix B. RTS is any study that had assessment = 1 or 2, F/U is assessment = 3, C/B is assessment = indeterminate/incomplete with best guess either unsure it exists, 2 or 3, and BX is assessment = indeterminate/incomplete with best guess either 4L, 4M, 4H or 5, or assessment = 4L, 4M, 4H or 5.

We also consider the binarization of these four categories into two groups: Normal and Not Normal. But there is controversy as to where the F/U category belongs, so we make its placement optional with either group. The point is to see if digitization or lossy compression make any difference to the fundamental decision made in screening: does the patient return to ordinary screening as normal, or is there suspicion of a problem and hence the demand for further work.

Truth is determined by agreement with a gold standard. The raw results are plotted as a collection of 2×2 tables, one for each category or group of categories of interest and for each radiologist. The differences among radiologists prove to be so large an effect that extreme care must be taken when doing any pooling or averaging of results across radiologists.

A typical table is shown in Table 3. The columns correspond to image modality or method I and the rows to II; I could be original analog and II original digitized, or I could be original digitized and II compressed digitized. "Right" and "Wrong" correspond to agreement or disagreement with the gold standard, respectively. The particular statistics could be, for example, the decision of "normal," i.e., return to ordinary screening. Regardless of statistic, the goal is to quantify the degree, if any, to which differences exist.

One way to quantify the existence of statistically significant differences is by an exact McNemar test, which is based on the following argument. If there are $N(1,2)$ entries in the (1,2) place and $N(2,1)$ in the (2,1) place, and the technologies are equal, then the conditional distribution of $N(1,2)$ given $N(1,2)+N(2,1)$ is binomial with parameters $N(1,2)+N(2,1)$ and 0.5; that is,

$$P(N(1,2) = k | N(1,2) + N(2,1) = n) = \left(\begin{array}{c} n \\ k \end{array} \right) 2^{-n};\ k = 0, 1, \ldots, n.$$

This is the conditional distribution under the null hypothesis that the two modalities are equivalent. The extent to which $N(1,2)$ differs from $(N(1,2)+N(2,1))/2$ is the extent to which the technologies were found to be different in the quality of performance with their use. Let $B(n, 1/2)$ denote a binomial random variable with these parameters. Then a statistically significant difference at level .05, say, will be detected if the observed k is so unlikely under the binomial distribution that a hypothesis test with size .05 would reject the null hypothesis if k were viewed. Thus if $\Pr(|B(n,1/2) - \frac{n}{2}| \geq |N(1,2) - \frac{n}{2}|) \leq .05$, then we declare a statistically significant difference has occurred.

Whether and how to agglomerate the multiple tables is an issue. Generally speaking, we stratify the data so that any test statistics we apply can be assumed to have sampling distributions that we could defend in practice. It is always interesting to simply pool the data within a radiologist

across all gold standard values, though it is really an analysis of the off-diagonal entries of such a table that is of primary interest. If we look at such a 4×4 table in advance of deciding upon which entry to focus, then we must contend with problems of multiple testing, which would lower the power of our various tests. Pooling the data within gold standard values but across radiologists is problematical because our radiologists are patently different in their clinical performances. This is consistent with what we found in an earlier study of MR and the measurement of the sizes of vessels in the chest [43, 44].

The counts can also be used to estimate a variety of interesting statistics, including sensitivity, predictive value positive (PVP) (also called positive predictive value or PPV), and specificity with respect to the personal and independent gold standards. An ROC-style curve can be produced by plotting the (sensitivity, specificity) pairs for the management decision for the levels of suspicion. Sample reuse methods (rather than common Gaussian assumptions) could be applied to provide confidence regions around the sample points [54].

A Wilcoxon signed rank test [55] can be employed to assess whether the subjective scores given to the analog originals, the uncompressed digitals, and the compressed images differ significantly from each other. With the Wilcoxon signed rank test, the significance of the difference between the bit rates is obtained by comparing a standardized value of the Wilcoxon statistic to two-tailed standard Gaussian probabilities. (The distribution of this standardized Wilcoxon is nearly Gaussian if the null hypothesis is true for samples as small as 20.) Our previous criticism of Gaussian assumptions are not relevant when they are applied to statistics for which the Central Limit Theorem is applicable.

Simple means and variances for the management statistics are presented in the Results section.

Learning Effects

The radiologists saw each study at least 5 times during the course of the entire experiment. These 5 versions were the analog originals, the digitized versions, and the 3 wavelet compressed versions. Some images would be seen more than 5 times, as there were JPEG compressed images, and there were also some repeated images, included in order to be able to directly measure intra-observer variability. We therefore needed to ascertain whether learning effects were significant. Learning and fatigue are both processes that might change the score of an image depending upon when it was seen.

In this work, we looked for whether learning effects were present in the management outcomes using what is known in statistics as a "runs" test [51]. We illustrate the method with an example. Suppose a study was seen exactly five times. The management outcomes take on four possible values (RTS, F/U, C/B, BX). Suppose that for a particular study and radiologist, the observed outcomes were BX three times and C/B two times. If there were no learning, then all possible "words" of length five with three BX's and two C/B's should be equally likely. There are 10 possible words that have three BX's and two C/B's. These words have the outcomes ordered by increasing session number; that is, in the chronological order in which they were produced. For these 10 words, we can count the number of times that a management outcome made on one version of a study differs from that made on the immediately previous version of the study. The number ranges from one (e.g., BX BX BX C/B C/B) to four (BX C/B BX C/B BX). The expected number of changes in management decision is 2.4, and the variance is 0.84. If the radiologists had learned from previous films, one would expect that there would be fewer changes of management prescription than would be seen by chance. This is a conditional runs test, which is to say that we are studying the conditional permutation distribution of the runs.

We assume that these "sequence data" are independent across studies for the fixed radiologist,

since examining films for one patient probably does not help in evaluating a different patient. So we can pool the studies by summing over studies the observed values of the number of changes, subtracting the summed (conditional) expected value, and dividing this by the square root of the sum of the (conditional) variances. The attained significance level (p-value) of the resultant Z value is the probability that a standard Gaussian is $\leq$ Z.

Those studies for which the management advice never changes have an observed number of changes 0. Such studies are not informative with regard to learning, since it is impossible to say whether unwavering management advice is the result of perfect learning that occurs with the very first version seen, or whether it is the result of the obvious alternative, that the study in question was clearly and independently the same each time, and the radiologist simply interpreted it the same way each time. Such studies, then, do not contribute in any way to the computation of the statistic. The JPEG versions and the repeated images, which are ignored in this analysis, are believed to make this analysis and p-values actually conservative. If no learning had occurred, then the additional versions make no difference. However, if learning did occur, then the additional versions (and additional learning) should mean that there would be even fewer management changes among the 5 versions that figure in this analysis.

Statistical Size and Power

Given a specified size, test statistic, null hypothesis, and alternative, statistical power can be estimated based using the common (but sometimes inappropriate) assumption that the data are Gaussian. As data are gathered, however, improved estimates can be obtained by modern computer intensive statistical methods. For example, the power and size can be computed for each test statistic described above to test the hypothesis that digital mammography of a specified bit rate is equal or superior to F/S mammography with the given statistic and alternative hypothesis to be suggested by the data. In the absence of data, we can only guess the behavior of the collected data to approximate the power and size. We consider a one-sided test with the "null hypothesis" that, whatever the criterion (management or detection sensitivity, specificity, or PVP), the digitally acquired mammograms or lossy compressed mammograms of a particular rate are worse than analog. The "alternative" is that they are better. In accordance with standard practice, we take our tests to have size .05. We here focus on sensitivity and specificity of management decisions, but the general approach can be extended to other tests and tasks.

Approximate computations of power devolve from the agreement tables of the form of Table 3, where the rows correspond to one technology (for example analog) and columns to the other (digital, say). The key idea is twofold. In the absence of data, a guess as to power can be computed using standard approximations. Once preliminary data are obtained, however, more accurate estimates can be obtained by simulation techniques taking advantage of the estimates inherent in the data. Table 4 shows the possibilities and their corresponding probabilities. The right hand column and bottom row are sums of what lies, respectively, to the left and above them. Thus, ψ is the value for one technology and $\psi + h$ is the value for the other; $h = 0$ denotes no difference. It is the null hypothesis. The four entries in the middle of the table are parameters that define probabilities for a single study. They are meant to be average values across radiologists, as are the sums that were cited. Our simulations allow for what we know to be the case: radiologists are very different in how they manage and in how they detect.

Two fundamental parameters are γ and R. The first is the chance (on average) that a radiologist is "wrong" for both technologies; R is the number of radiologists. These key parameters can be estimated from the counts of Table 3 resulting from the pilot experiment, and then improved as

additional data is acquired.

Prevalence

The pilot data set of 57 images has two obvious shortcomings: it is too small to have good power for tests of reasonable size for those tests proposed, and the prevalence of abnormalities in this data set does not accurately reflect that of a normal screening population. This violates the literal goals of accurate simulation and representative statistics for a screening application. The first shortcoming can be resolved by a larger study, although it is a serious and controversial issue as to how large the study must be. The second problem, however, is unavoidable with any study of reasonable sample size as prevalence in a screening population can vary widely at different locations. We argue, however, that relevant conclusions can be drawn for the true prevalence based on a carefully constructed study using different proportions. In order to well simulate the proportion of normal images to ones containing pathology that actually would be found in a screening situation, we would require thousands of studies as there are only 6–8 cancers/1000 asymptomatic women screened. In our approach we do not *directly* estimate overall statistics for detection (sensitivity, PVP) and management (sensitivity, specificity). This would result in little power for some of the statistics without unreasonably large patient numbers or unreasonably large size to the tests. It would also involve incorporating somewhat arbitrarily prevalence values for the abnormalities. A purely prospective screening study using commonly assumed prevalence values could require more than 11,000 patients. Our "retrospective/prospective" approach, [50], allows us to compute estimates of our statistics conditional on the presence or absence of abnormalities and to estimate separately size and power for both conditional populations. This then yields by straightforward algebra overall statistics by suitably weighting the conditional statistics to reflect estimated prevalence. The specific numbers of patients needed for good size and power will be estimated in a cumulatively improving manner as the data are gathered and the experiments performed. Preliminary analysis based on standard approximations suggests that this is on the order of $400 - 600$.

One reasonable concern about not attempting to simulate accurately a population prevalence is that radiologists might behave differently if they knew that the prevalences in an experiment were different from that ordinarily encountered in a clinic. This effect could be analyzed in a quantifiable manner by varying the prevalence at different sites in a controlled manner not known to the judges or assistants.

Lesion Detection Accuracy

The question of lesion detection accuracy is of comparable importance to the issue of management. For the digital technologies to be useful, true lesions detectable in the analog images should be detectable in the digital. This section explores the theoretical basis for statistical analysis of the lesion detection problem.

The first issue that needs to be explored is the notion of two lesions being declared equivalent. We first define what is meant by a lesion (also referred to as a *finding*). For the purposes of detection, we will treat a finding as being a dominant observation made by a radiologist *belonging to a single view.* This differs slightly from the use of the word elsewhere in this report, where for example an individual finding might be seen in both the MLO and CC views of a patient. For detection purposes we consider this to be two separate findings (albeit separate findings sharing much in common, such as their management category, etc.). Thus, with some oversimplification, a finding for detection purposes may be thought of as a single circle marked by a radiologist on a transparency. The rational for so doing is best explained by way of an example. Suppose a doctor

has found a abnormality in both the MLO and CC views of a patient when viewed at 0.15 bpp. The same doctor detects the abnormality in only the MLO view when presented with the analog images of the same patient. An expert radiologist, on examining the diagnoses, declares that the doctor was referring to the same abnormality but for whatever reason only saw it in the MLO view on the analog films. Do we declare these findings, for the purposes of analysis, equivalent or not? To declare them equivalent would bias statistics towards declaring the technologies equivalent. To declare them different would bias results in the opposite direction. Instead we treat the 0.15 bpp finding as actually being two separate entities. For analysis we would declare that we have one matched finding and one unmatched finding, thus achieving a compromise between the two extremes illustrated in the example.

We now explore in greater depth the notion of two findings being declared equivalent. For any two findings x and y let us define a function $\mathcal{D}$ that equals 1 if the findings are equivalent and zero otherwise. A finding has associated with it a great deal of information: the view it was visible in, its physical location in the image, its size, its type, its management category, the name of the doctor who found it, etc. The function $\mathcal{D}$ has available to it all of this information plus whatever rules are built into it. For example, if $\mathcal{D}$ is the pronouncement of an expert radiologist, all of the radiologist's training factors into the decision function. Let us refer to the function used by an expert radiologists as $\mathcal{D}^*$ and declare it to be truth. We seek a function $\mathcal{D}^\dagger$ that can be computed on a digital computer and mimics $\mathcal{D}^*$ as closely as possible. Hereafter, when we refer to two findings x and y being equivalent we mean that $\mathcal{D}^\dagger(x, y) = 1$ and conversely.

In this experiment, the rules used by $\mathcal{D}^\dagger$ were purely geometrical ones. Findings were declared equivalent if they were seen in the same view (of a particular patient) and physically overlapped. This was a computable function implementable given the data collected in the experiment. It was also found to be in good agreement with $\mathcal{D}^*$, although we do not have quantitative measures of the agreement at this time.

Given both $\mathcal{D}^\dagger$ and a gold standard, a value can be assigned to the sensitivity, the probability that something is detected given that it is present in the gold standard. A judge who labels abnormalities everywhere in an image could have perfect sensitivity. Predictive value positive, the chance an abnormality is actually present given that it is marked, fills the role of specificity in penalizing false positive reporting. A judge who is too aggressive in finding abnormality could have high sensitivity at the expense of low PVP while a judge who is too stringent about what defines abnormality could have a high PVP at the expense of low sensitivity. As is the case with the ROC parameters of true positives and false positives, both sensitivity and PVP will be 1 if the decision is perfect.

Given that we can calculate sensitivity and PVP values at different bit rates, we need a method of comparing them across these rates. The comparison used in this experiment was carried out using a permutation distribution of a two–sample t–test that is sometimes called the Behrens–Fisher test [52, 53]. The statistic takes account of the fact that the within group variances are different. The test is exact and does not rely on Gaussian assumptions that would be patently false for this data set. The use of this statistic is illustrated by the following example. Suppose a judge has judged N patients at both rates A and B (where a rate includes the analog original images). The images can be divided into m groups according to whether the gold standard for the image contained $0, 1, \ldots, m$ findings. Let N_i be the number of images in the ith group. Let Δ_{ij} represent the difference in sensitivities (or PVP) for the jth image in the ith group seen at rate A and at rate B. Let $\overline{\Delta}_i$ be the average difference:

$$\overline{\Delta}_i = \frac{1}{N_i} \sum_j \Delta_{ij}.$$

We define

$$S_i^2 = \frac{1}{N_i - 1} \sum_j (\Delta_{ij} - \overline{\Delta}_i)^2$$

and then the Behrens–Fisher t statistic is given by

$$t_{BF} = \frac{\sum_i \overline{\Delta}_i}{\sqrt{\sum_i S_i^2 / N_i}}.$$

The Δ_{ij} are fractions with denominators not more than m so they are are utterly non–Gaussian. Therefore, computations of attained significance (p values) are based on the restricted permutation distribution of t_{BF}. For each of the N patients, we can permute the results from the two rates or not. There are 2^N points possible in the full permutation distribution and we calculate t_{BF} for each one. The motivation for the permutation distribution is that if there were no differences between the rates, then in computing the differences Δ_{ij} it should not matter whether we compute rate A minus rate B or vice versa. We would not expect the "real" t_{BF} to be an extreme value amount the 2^N values. If k is the number of permuted t_{BF} values that exceed the "real" one then $(k+1)/2^N$ is the attained one–sided significance level for the test of the null hypothesis that the lower rate performs at least as well as the higher one. The one–sided test of significance is chosen to be conservative and to argue most strongly against compression.

4 Compression Algorithms

We use a compression algorithm of the subband/pyramid/wavelet coding class. These codes typically decompose the image using an octave subband, critically sampled pyramid, or complete wavelet transformation, and then code the resulting transform coefficients in an efficient way. The decomposition is typically produced by an analysis filter bank followed by downsampling. Any or all of the resulting subbands can be further input to an analysis filter bank and downsampling operation, for as many stages as desired. The most efficient wavelet coding techniques exploit both the spatial and frequency localization of wavelets. The idea is to group coefficients of comparable significance across scales by spatial location in bands oriented in the same direction. The early approach of Lewis and Knowles [25] was extended by Shapiro in his landmark paper on embedded zerotree wavelet coding [27] and the best performing schemes are descendents or variations on this theme. The approach provides codes with excellent rate-distortion tradeoffs, modest implementation complexity, and an embedded bit stream, which makes the codes useful for applications where scalability or progressive coding are important. Scalability implies there is a "successive approximation" property in the bit stream. As the decoder gets more bits from the encoder, the decoder can decode a progressively better reconstruction of the image. This feature is particularly attractive for a number of applications, especially those where one wishes to view an image as soon as bits begin to arrive, and where the image improves as further bits accumulate. With scalable coding, a single encoder can provide a variety of rates to customers with different channels or display capabilities. Since images can be reconstructed to increasing quality as additional bits arrive, it provides a natural means of adjusting to changing channel capacities and a more effective means of using a relatively slow channel. This approach provides compression algorithms that are among the very best available in terms of quality for a given bit rate and they are naturally progressive, scalable, and of modest computational complexity.

After experimenting with a variety of algorithms, we chose Said and Pearlman's variation [28, 29] of Shapiro's embedded zerotree algorithm [27] because of its good performance and the availability

of working software for 12 bpp originals. We use the default filters (the 9-7 biorthogonal filters) in the software compression package of Said and Pearlman [28]. These filters are considered, for example, in Antonini [23] and Villasenor et al. [58]. A description and discussion of the algorithm along with access to the software may be found at the World Wide Web site [29]. The algorithm applies a succession of thresholds to each coefficient, each half the size of the preceding. Coefficients with magnitude smaller than the threshold are deemed insignificant and are effectively quantized to zero. Bits are sent only to indicate the location of pixels that fall above the thresholds, and they are sent in an order determined by a subset partitioning algorithm that takes advantage of the correlation across scales of significance according to spatial location and orientation. Once a pixel is deemed significant, further bits sent regarding that pixel are devoted to refining the accuracy of the actual location by bit plane transmission. The bits are sent so as to first describe the largest coefficients, which contribute the most to the reconstruction accuracy. In this way the bit stream can be stopped at any point with a good reproduction for the given number of bits. The system incorporates the adaptive arithmetic coding algorithm considered in Witten, Neal, and Cleary [59].

For our experiment additional compression was achieved by a simple segmentation of the image using a thresholding rule. This segmented the image into a rectangular portion containing the breast — the *region of interest* or *ROI* — and a background portion containing the dark area and any alphanumeric data. The background/label portion of the image was coded using the same algorithm, but at only 0.07 bpp, resulting in higher distortion there. We here report SNRs and bit rates for both the full image and for the ROI.

The image test set was compressed in this manner to three bit rates: 1.75 bpp, 0.4 bpp, and 0.15 bpp, where the bit rates refer to rates in ROI. The average bit rates for the full image thus depended on the size of the ROI. An example of the Said-Pearlman algorithm with a 12 bpp original and 0.15 bpp reproduction is given in Figure 1. For comparison purposes we also compressed a few images using JPEG [18]. In the JPEG example the background and labels were coded by using JPEG with default quantization table settings so that the bit rates are not directly comparable.

We also investigated new compression algorithms, including two new forms of multiresolution vector quantization algorithms. These included vector generalizations of the embedded zerotree wavelet technique [60, 61] and a non-wavelet multiresolution technique using tree-structured codes [62]. The first method used a variable-rate tree-structured vector quantizer applied to the coefficients produced by an orthogonal wavelet decomposition. The set of vectors from different levels of the decomposition that correspond to the same orientation and spatial location were examined in various zerotree groups to determine the different bit rates and distortions achievable for the set. The decision not to code certain groups of vectors was based upon choosing the desired distortion/rate tradeoff from among the possibilities. Side information was sent to the decoder to inform it of the sequence of decisions. The resulting bit stream was entropy coded. Results of this method yielded a peak signal-to-noise ratio of 30.16 dB at 0.148 bpp on a test image, a slight improvement on the scalar zerotree codes. The incurred additional complexity, however, made this approach inferior to the scalar wavelet coding scheme selected.

The second method was an approach to multiresolution image coding without using a wavelet decomposition. Here a multiresolution tree structured vector quantizer was developed to produce an embedded code, so that the quality of the image is optimized for the corresponding resolution at any number of bits. The resolution at which the image is viewed given a particular number of bits is determined by the specific decoder. The multiresolution tree structured vector quantizer generates the codebook by greedy tree growing, which is an extension of the generalized Breiman-Friedman-Olshen-Stone BFOS algorithm [63, 70]. The tree is grown one step further by splitting the node which will yield the best ratio of the change in distortion at the corresponding resolution

of current bit rate to the change in rate. The decoder has codewords of all resolutions obtained by optimal centroiding for a given resolution and a given encoder partition. The encoding of an image is essentially the same as BFOS algorithm and the difference is that instead of having a fixed distortion measure, the distortion measure is defined for the corresponding resolution at a particular bit rate. The results provided significant improvements at modest complexity for low resolution images, e.g., for medical images reduced in size for progressive viewing during the rendering of the full image. At the full size required for screening or diagnostic viewing, however, the quality was not comparable to the compression algorithm adopted. They may prove useful, however, in teleradiology applications as they can provide interim high quality small images which may in some cases speed ROI selection or identification of problems.

5 Compression and Classification

Classification and compression both consist of the mapping of real valued vectors (such as pixel intensity blocks in an image) into a finite set. If the goal is classification, the set is a collection of classes such as tumor and nontumor. However, if the goal is compression, the set is a collection of templates or reproduction codewords. The history of these two fields is intertwined and many similar algorithms have been developed for the two applications, such as the CART [63] (classification and regression tree) algorithm for classification trees and tree-structured vector quantization for compression. Both are designed by optimally trading off a measure of quality, such as mean squared error for compression or Bayes risk for classification, with a measure of rate or complexity, such as the average number of bits required to make a decision or represent the template.

By incorporating a Bayes risk term into the usual distortion measure for a compression system, one can simultaneously optimize a code with respect to compression, classification or any weighted combination of the two. This is done within the framework of vector quantization (VQ), since there exists an intimate connection between the algorithms used to design and implement vector quantizers for compression and those for statistical classification that provides a natural means of jointly optimizing the two goals. The input is a joint random process $\{X(n), Y(n); n = 0, 1, \ldots\}$, where the $X(n)$ are k-dimensional real-valued vectors (pixel blocks in an image compression application) and the $Y(n)$ designate membership in a class and take values in a set $\mathcal{H} = \{0, 1, \cdots, M-1\}$. The VQ-based classifier operates solely on the observed sequence X and consists of an encoder α, which views only X and outputs a binary index $i = \alpha(X) \in \mathcal{I}$ and a decoder β, which maps the indices into the reproduction vectors $\beta(i) = \hat{X}_i$ and a class label $\delta(i) \in \mathcal{H}$. Because the index i is used to both decompress and classify the vector, the classification is implicit in the compression, and hence no additional computation or bits are required.

The quality of the reproduction $\hat{X} = \beta(\alpha(X))$ for an input X is measured by a nonnegative distortion $d(X, \hat{X})$, typically taken to be the squared error distortion, $d(X, \hat{X}) = ||X - \hat{X}||^2$, for simplicity. The average distortion for compression $D(\alpha, \beta) = E[d(X, \beta(\alpha(X)))]$ is then the mean squared error (MSE). The quality of the classifier is measured by the Bayes risk, defined as $B(\alpha, \delta) = \sum_{i=0}^{N-1} P(\alpha(X) = i) \times \sum_{j=0}^{M-1} C_{j,\delta(i)} P(Y = j | \alpha(X) = i)$ where the cost $C_{j,k} \geq 0; k = \delta(i)$ represents the cost incurred when a class j vector is classified as class k, where $C_{j,k} = 0$ if $j = k = \delta(i)$. These costs can be chosen to reflect the fact that different classification errors can have different consequences, as do the presence or absence of tumors. If the nonzero costs are chosen to be equal, then the Bayes risk reduces to the probability of classification error.

The Bayes VQ system [64, 65, 66, 67, 68, 69] consists of two cascaded vector quantizers. The design of these vector quantizers is based on the empirical distribution $P_{\mathcal{L}}$ induced by a training set of labeled data $\mathcal{L} = \{(x_n, y_n); n = 1, \ldots, |\mathcal{L}|\}$. The first stage in the system, a tree-structured

vector quantizer (TSVQ), provides an estimate of the posterior conditional probabilities required to compute the Bayes risk weighted distortion. It is designed in a manner analogous to the CART algorithm to generate estimates of $P(Y = l|X = x) = P_{\mathcal{L}}(l|x)$. The TSVQ is designed to minimize the distortion given by the average relative entropy $D(P_{\mathcal{L}}||\hat{P}) = \sum_{x\in\mathcal{L}} P_{\mathcal{L}}(x) \sum_{l\in\mathcal{H}} P_{\mathcal{L}}(l|x) \log \frac{P_{\mathcal{L}}(l|x)}{\hat{P}(l|x)}$ between the estimated probability and the empirical probability. Other methods of generating these necessary probability estimates are under investigation in another project and will eventually be tested on the mammographic data. Squared error determines which node is used to encode a vector within the tree. The probability estimate for any node is then given by the relative frequencies of the class labels given the encoder output. These encoded estimates are used in creating the second VQ and are not communicated to the decoder.

The second VQ, either full search or tree-structured, incorporates the simultaneous optimization of both compression and classification by using a modified distortion measure that contains both squared error for general appearance and Bayes risk for classification accuracy. These two error measures are combined with a Lagrangian importance weighting to form the modified distortion measure

$$\rho_{\lambda,\hat{P}}(x, \hat{x}, l) = ||x - \hat{x}||^2 + \lambda \sum_{j=0}^{M-1} C_{j,l} \hat{P}(Y = l|x)$$

so that

$$\begin{aligned} J_{\lambda,\hat{P}}(\alpha, \beta, \delta) &= E[\rho_{\lambda,\hat{P}}(X, \beta(\alpha(X)), \delta(\alpha(X)))] \\ &= D(\alpha, \beta) + \lambda B(\alpha, \gamma). \end{aligned}$$

This VQ is designed using a descent algorithm, analogous to the Lloyd clustering algorithm algorithm [70], that iteratively optimizes the encoder, decoder, and classifier for each other. The classifier is a minimum average Bayes risk classifier, defined by $\delta_{\text{Bayes}}(i) = \arg\min_k \sum_{j=0}^{M-1} C_{j,k} \hat{P}(Y = j|\alpha(x))$. The costs, $C_{j,k}$, are particularly important in the classification of mammograms since the consequences of misclassifying an abnormality (i.e. missing a tumor) are quite different than those of a false alarm. The Lagrangian parameter λ provides a flexible tradeoff between compression and classification priorities. In particular, when $\lambda = 0$, the focus for designing the encoder and decoder is purely on compression; we thus create an ordinary minimum MSE VQ. If a class label, optimized for the VQ encoder output, is subsequently attached to these vectors, the system is simply an independent design of a VQ and a classifier. When $\lambda \to \infty$, we obtain a minimum Bayes risk classifier.

6 Results and Discussion

6.1 Clinical Experiment

The clinical experiment took place in two phases, at Stanford University Hospital during spring 1996 and at the University of Virginia during summer 1996. The experiment required roughly thirty hours of time for each radiologist (donated) and 500 hours of student assistant or technician time to schedule and run the sessions, hang overlays and films, and enter the data. The doctors participating in the study at Stanford were R. Birdwell, M.D., S. Rossiter, M.D. and B.L. Daniel, M.D. The University of Virginia doctors were L.J. Fajardo, M.D., R. Moran, M.D., and G. DeAngelis, M.D. The gold standard was established by E. Sickles, M.D., Professor of Radiology, University

of California at San Francisco, and Chief of Radiology, Mt. Zion Hospital, and D. Ikeda, M.D., Assistant Professor and Chief, Breast Imaging Section, Dept. of Radiology, Stanford University, an independent panel of expert radiologists, who evaluated the test cases and then collaborated to reach agreement. The majority of the detected items were seen by both radiologists. All findings were included, even if seen by only one radiologist. The other type of discrepancy resolved was the level of suspicion of the detected lesions. Since the same abnormality may be classified differently, the two radiologists were asked to agree.

6.2 SNR vs. Bit Rate

The SNRs are summarized in Tables 6 and 7. The SNR definition is $10 \log_{10} \mathrm{E/MSE}$, where MSE denotes the average squared error and E denotes the energy of the digital original pixels. The overall averages are reported as well as the averages for the specific image types or views (left and right breast, CC and MLO view). This demonstrates the variability among various image types as well as the overall performance. Two sets of SNRs and bit rates are reported: ROI only and full image. For the ROI SNR the rates are identical and correspond to the nominal rate of the code used in the ROI. For the full images the rates vary since the ROI code is used in one portion of the image and a much lower rate code is used in the remaining background and the average depends on the size of the ROI, which varies among the images. A scatter plot of the ROI SNRs is presented in Figure 2, where each image in the study provides a point.

For comparison purposes we report the results for optimized JPEG for the ROI in Table 8. The primary observation is that the numbers in the corresponding positions of the wavelet coding table and the images are visually similar, but the rates needed to make JPEG comparable to the wavelet scheme are more than 1 bpp higher, a significant amount at the lower bit rates. The bit rates listed were target bit rates and the actual bit rates achieved varied slightly.

6.3 Management Differences

The focus of this section is on the screening and management of patients and how they are is affected by analog vs. digital and lossy compressed digital. In all, there were 57 studies. According to the gold standard, the respective numbers of studies of each of the four management types RTS, F/U, C/B, and BX were 12, 1, 18, and 26, respectively.

For each of the four possible outcomes, the analog original is compared to each of four technologies: digitized from analog original, and wavelet compressed to three different levels of compression (1.75 bpp, 0.4 bpp, and 0.15 bpp). The McNemar 2×2 statistics based on the generic table of Table 3 for assessing differences between technologies were computed 48 times, 16 per radiologist, for each competing image modality (original digital and the three lossy compressed bit rates). For example, the 2×2 tables for a single radiologist comparing analog to each of the other four modalities is shown in Table 9 (the three Stanford radiologists are referred to as A, B and C, the three UVa radiologists as D,E and F). For none of these tables for any radiologist was the exact binomial attained significance level (p-value) .05 or less. For our study and for this analysis, there is nothing to choose in terms of being "better" among the analog original, its digitized version, and three levels of compression, one rather extreme. We admit freely that this limited study had insufficient power to permit us to detect small differences in management. The larger the putative difference, the better our power to have detected it.

Tables 10 and 11 summarize the performance of each radiologist on the analog vs. uncompressed digital and lossy compressed digital images. In all cases, columns are "digital" and rows "analog".

Were it the case that a doctor's management decisions were identical at two different rates, all off diagonal entries would be zero. Table 10(A) treats analog vs. original digital and Tables 10(B)–(D) treat analog vs. lossy compressed digital at bit rates of 1.75 bpp, 0.4 bpp, and 0.15 bpp, respectively. Consider, for example, Table 10(D) which compares the analog images with the most compressed images for radiologist (A). By summing up the fourth row we see that (A) sent 18 patients to biopsy (the most critical management category) when viewing analog images. By summing the fourth column, we see that the same doctor sent 20 patients to biopsy when viewed at 0.15 bpp.

Tables 10 and 11 suggest that radiologists differ substantially from each other. For example, radiologist (B) is highly likely to biopsy a patient independent of modality, as can be observed by examination of the $(4, 4)$ entries of this doctor's matrices. Radiologist (E), on the other hand, was less likely to send a patient to biopsy, again independent of modality. Comparing radiologists was not a goal of this study; we are interested in what happens when a particular radiologist views the same image under different modalities. The differences among radiologists merely make it more difficult to evaluate the differences among analog, digital, and lossy compressed images, since extreme care must be taken when doing any pooling or averaging of results across radiologists. Nonetheless, a primary conclusion from the data and analysis is that variabilities among judges exceed by a considerable amount, in their main effects and interactions, the variability in performance that owes to imaging modality or compression within very broad limits. In other words, the differences among analog, digital, and lossy compressed images are in the noise of the differences among radiologists, and are therefore more difficult to evaluate.

The runs test for learning did not find any learning effect at the 5% significance level for these management outcomes. For each of the 3 judges, approximately half of the studies were not included in the computation of the statistic, since the management decision was unchanging. For the 3 judges, the numbers of studies retained in the computation were 28, 28, and 27. The Z values obtained were -0.12, -0.86, and -0.22, with corresponding p-values of 0.452, 0.195, and 0.413. Further testing for learning will include an analysis of the detected findings.

Management Sensitivity and Specificity

The means and variances of the sensitivity and specificity and the mean of the PVP of the management decisions with respect to the independent gold standard are summarized in Table 12. In this table, sensitivity, specificity, and PVP are defined relative to the independent gold standard. The table does not show any obvious trends for these parameters as a function of bit rate. Sensitivity is the ratio of the number of cases a judge calls "positive" to the number of cases actually "positive" according to the independent gold standard. Here "positive" is defined as the union of categories F/U, C/B, and BX. A "negative" study is RTS. Sensitivity and specificity can be thought of as binomial issues, and so if the sensitivity is p, then the variance associated with that sensitivity is $p(1-p)$. Methods are being developed to provide joint bootstrapped confidence intervals for sensitivity and specificity. The standard deviation calculation for PVP is somewhat more complicated and is not included here; because PVP is the ratio of two random quantities (even given the gold standard), the variance calculation requires approximate statistical methods as in analyses by "propagation of errors."

Size and Power

In our small pilot study of management, we found sensitivity of about .60 and specificity about .55. The respective estimated values of h varied from more than .02 to about .07; γ was about

.05. These numbers are all corrupted by substantial noise. Indeed, the variability associated with our estimation of them is swamped by the evident variability among radiologists. For a test of size .05, by varying parameters in amounts like what we saw, the power might be as low as .17 with 18 radiologists, or as high as 1.00 with only 9 radiologists. The power is very sensitive to the three parameters, and there are not yet adequate data to have make a precise estimate. No matter how many studies or how many radiologists we would have, one could always vary the parameters so that we would need more of either or both.

If we think of sensitivity for detection being .85, say, then at least for that quantity 400 studies and 9 radiologists seem ample. At this time the best recommendation would be to start with the 400 studies we have recommended in the past, 12 radiologists, three at each of four centers, and find an attained significance level for a test of the null hypothesis that there is no difference between technologies. And, perhaps at least as important, estimate the parameters of Table 4. At that point possible numbers of required further radiologists or studies, if any, could be estimated for particular values of size and power that reviewers might require. The design could be varied so that the pool of studies would include more than 400, but no single radiologist would read more than 400. In this way we could assess fairly easily the impact of variable prevalence of adverse findings in the gold standard, though we could get at that issue even in the situation we study here.

Computations of power apply equally well in our formulation to sensitivity and specificity. They are based on a sample of 400 studies for which prudent medical practice would dictate RTS for 200, and something else (F/U, C/B, or BX) for the other 200. Thus, there are 200 studies that figure in computation of sensitivity and the same number for specificity. All comparisons are in the context of "clinical management," which can be "right" or "wrong." It is a given that there is an agreed upon gold standard, independent or separate. For a given radiologist who has judged two technologies – here called I and II and meant to be digital and analog or analog and lossy compressed digital in application – a particular study leads to an entry in a table of the form Table 3. Table 5 summarizes the probability estimates formed by dividing the counts of Table 3 by N, the number of studies which are not normal.

If the null hypothesis of "no difference in technologies" is true, then whatever be ψ and γ, $h = 0$. An alternative hypothesis would specify $h \neq 0$, and without loss (since we are free to call whichever technology we want I or II) we may take $h > 0$ under the alternative hypothesis that there is a true difference in technologies. Under the null, *given* $b + c$, b has a binomial distribution with parameters $b + c$ and $1/2$. Under the alternative, given $b + c$, b is binomial with parameters $b + c$ and $(1-\psi-\mathrm{h}-\gamma)/(2-2\psi-2\gamma-\mathrm{h})$. The usual McNemar *conditional* test of the null hypothesis is based on $(b-c)^2/(b+c)$ having approximately a chi-square distribution with one degree of freedom.

Had the project continued, we would have studied the use of 9 or more radiologists by assuming that their findings were independent and combining their data by adding the respective values of their McNemar statistics. We always intend that the *size* = probability of Type I error being .05. Since the sum of independent chi-square random variables is distributed as chi-square with degrees of freedom the sum of degrees of the respective degrees of freedom, it is appropriate to take as the critical value for our test the number C, where $\Pr(\chi^2_R > C) = .05$. The four respective values of C are therefore 16.92, 21.03, 25.00, and 28.87.

Computation of power is tricky because it is *unconditional* since before the experiment, $b+c$ for each radiologist is random. Thus, the power is the probability that a non-central chi-square random variable with R degrees of freedom and non-centrality parameter $[(p_1-.5)^2/p_1q_1]\sum_{i=1}^{R}(b_i+c_i)$ exceeds $C/4p_1q_1$, where b_i+c_i has a binomial distribution with parameters N and $2-2\psi-2\gamma-h$; and the R random integers are independent; $p_1 = (1-\psi-h-\gamma)/(2-2\psi-2\gamma-h) = 1-q_1$. This entails that the non-centrality parameters of the chi-square random variable that figures in the computation

of power is itself random. Note that a non-central chi-square random variable with R degrees of freedom and non-centrality parameter Q is the distribution of $(G_1 + Q^{1/2})^2 + G_2^2 + \cdots + G_R^2$, where $G_1, ...G_R$ are independent, identically distributed standard Gaussians. On the basis of previous work and pilot study, we have chosen to compute the power of our size .05 tests for N always 200, ψ from .55 to .85 in increments of .05; $\gamma = .03, .05, .10$; and, as was stated, $R = 9, 12, 15$, and 18. The simulated values of power are presented in Table 13. These form the basis of our earlier estimates for the necessary number of patients and should be updated as data is acquired.

6.4 Lesion Detection

The analysis of the lesion detection problem is not yet complete. The final analysis will be submitted at a future date. The results obtained so far have supported the central hypothesis that there is no difference between the imaging modalities from a lesion detection viewpoint. This section presents the results obtained so far, draws conclusions from them, and indicates what analysis still needs to be done.

The data subset which has been analyzed consists of those observations recorded by the UVa judges. Sensitivity and PVP values were obtained for each judge with respect to a personal gold standard defined by their analog original viewings. The data are summarized in Table 14. The results of the Behrens–Fisher t–statistic indicated no statistically significant differences at the 5% significance level for either the sensitivity or PVP results. Indeed, there were no differences at the 10% significance level. The results therefore support the central hypothesis.

The use of a personal gold standard, as explained elsewhere in this report, is not desirable. By its use we can at best show that the digital images are equivalent to the analog images. It would have been preferable to use the independent gold standard, as was done for the management analysis. Unfortunately, neither the independent gold standard nor the Stanford data subset was ready for lesion analysis as of the report deadline. The reasons for the delay are discussed later in the report but essentially involved a shortage of personnel to perform laborious retroactive measurements. The process of bringing the independent gold standard and the Stanford data set to a level whereby they can have lesion analysis performed on them is underway. Once completed, the lesion analysis can be repeated with respect to the consensus gold standard on a larger data set.

6.5 Subjective Ratings vs. Bit Rate

In the previous sections, objective measures of the quality of the compressed images were analyzed via the SNR values and patient management decisions on the digitally compressed images. It is also informative to examine the effects of compression on subjective opinions. Table 15 provides the means and standard deviations for the subjective scores for each radiologist separately and for the radiologists pooled. The distribution of these subjective scores are displayed in Figures 3–5.

Figure 3 displays the frequency for each of the subjective scores obtained with the analog images. Figure 4 displays the frequency for each of the subjective scores obtained with the uncompressed digital images (judges pooled), and Figure 5 displays the frequency for each of the subjective scores obtained with the digital images at Level 3.

Using the Wilcoxon signed rank test, the results were as follows:

Judge A: All levels were significantly different from each other except the digital to .4 bpp, digital to 1.75 bpp, and .4 to 1.75 bpp.

Judge B: The only differences that were significant were .15 bpp to .4 bpp and .15 bpp to digital.

Judge C: All differences significant .

All judges pooled: All differences were significant except digital to .15 bpp, digital to 1.75 bpp, .15 to .4 bpp, and .15 to 1.75 bpp.

Comparing differences from the independent gold standard, for Judge A all were significant except digital uncompressed, for Judge B all were significant, and for Judge C all were significant except 1.75 bpp. When the judges were pooled, all differences were significant.

There were many statistically significant differences in subjective ratings between the analog and the various digital modalities, but some of these may have been a result of the different printing processes used to create the original analog films and the films printed from digital files. The films were clearly different in size and in background intensity. The judges in particular expressed dissatisfaction with the fact that the background in the digitally produced films was not as dark as that of the photographic films, even though this ideally had nothing to do with their diagnostic and management decisions.

6.6 Combined Compression and Classification

For this experiment a set of 12 bit 100 micron resolution digitized mammograms was used. The images contained calcifications or masses. Two different training sets were formed. One training set consisted of mammograms with calcifications and the other training set contained mammograms with masses. Each training vector consisted of a 2x2 pixel block of intensity values and a class label. We considered two test images taken from outside of the training sets. One test image contained calcifications and was encoded using the calcification codebook. The other test image contained masses and was encoded using the mass codebook.

SNR was measured as $10 \log_{10}(D_0/D)$, D being the distortion measured by mean squared error, and D_0 the distortion obtained on the test sequence using the best zero rate code. The classification performance is measured by three parameters: Bayes risk, sensitivity and specificity.

We used a tree-structured compression code for the Bayes VQ system (BTSVQ design). We compare our results with an independent tree-structured design of classifier and quantizer (independent TSVQ design), corresponding to the special case of $\lambda = 0$ in a Bayes VQ system. We also compare the performance of our system to Kohonen's "learning vector quantizer" (LVQ) [71], a popular classification method for a variety of applications. LVQ implicitly designs a full search codebook to reduce classification error rather than reducing compression error. The encoder operates as an ordinary minimum squared error selection of a representative from the codebook. Because the algorithm does not explicitly consider compression in its design, we use the codebook only for classification purposes. An optimized version of this codebook, which replaces encoder codewords produced by LVQ by the centroids of all training vectors that map into them, is then used to provide the reproduction vectors [67]. For the LVQ design, the codebook was initialized using the LVQ PAK *propinit* algorithm and then designed using the *olvq1* with the modification discussed above. We note that LVQ does not have the capability to incorporate unbalanced misclassification costs into its design.

For the analysis of the mammogram with calcifications, we designated the cost of misclassifying a tumor vector as 50 times more detrimental than misclassifying a nontumor vector. We selected $\lambda = 10^5$ for the BTSVQ design. Figure 6 presents results obtained on a portion of the test image containing calcifications at 2bpp.

The table indicates that the BTSVQ design produced lower Bayes risk and comparable SNR to the independent TSVQ design for the test image containing calcifications. The table also indicates

that the BTSVQ design significantly outperformed LVQ with respect to both SNR and Bayes risk. The BTSVQ design yielded higher specificity but lower sensitivity to that obtained with the independent TSVQ design. Because of the preponderance of nontumor vectors, the 3.15% difference in specificity can affect the visual quality of the images considerably. This is illustrated in Figure 1, where we observe less false highlighting with the BTSVQ design than with the independent TSVQ design. Although the sensitivity obtained for the encoded calcification image using the Bayes TSVQ method seems low, if we use a criterion that is less stringent than the *pixel-by-pixel* definition that asks whether the algorithm has detected enough of the lesion to signal the radiologist, the answer to such a question would be affirmative.

For the analysis of the mammogram with a mass, we designated the cost of misclassifying a mass vector as ten times more harmful than a false alarm. The BTSVQ design again used $\lambda = 10^5$. Table 17 provides the statistical results obtained using the BTSVQ and independent TSVQ designs on the test image containing masses at 2 bpp. LVQ did poorly on this image (none of the mass vectors were classified correctly and its compression performance was very low), and as such a comparison with LVQ here is not meaningful.

We again observe that the the BTSVQ design produced lower Bayes risk and comparable SNR to the independent TSVQ design. In addition, the BTSVQ design yielded higher specificity but lower sensitivity to that obtained with the independent design. The BTSVQ design correctly identified the one mass in the corresponding test image section including its spiculation. The independent design, however, did not clearly point out the mass area. Although both algorithms produced false highlighting, the amount engendered by the BTSVQ design was considerably less than that of the independent design (a 5.1% difference).

For both calcification and mass images, the BTSVQ and independent designs produced a tradeoff between specificity and sensitivity, i.e., an increase in performance in one of the measures resulted in a decrease in performance in the other. We note, however, that only the Bayes TSVQ design has the ability to select a desirable ratio between these two measures using both costs and the λ parameter.

The BTSVQ design produced lower Bayes risk and visually superior compressed/classified images compared to those obtained with the independent TSVQ design and with LVQ on the two test images we investigated. Although, the general area of the lesions were identified by the BTSVQ algorithm, we obtained some false highlighting as well – particularly with the mammograms containing masses. We note, however, that the λ value and costs inherent in the BTSVQ design allow a flexibility in the compression and classification performance.

6.7 Observations on Implementation

In the course of performing the experiment several observations and conclusions were drawn concerning the details of implementation. As the experiment proceeded, certain items were fine tuned so as to assist eventual statistical analysis or to save human labor. This section of the report highlights a few of the more important observations and conclusions. Some are related to the actual process of collecting data, others to the eventual analysis of the data.

Data Collection

The data collection process was a long and laborious one, involving many hours on the part of both radiologists and student assistants. Details of the experimental implementation have already been discussed and are elaborated upon in the appendix. Worth highlighting is the importance of

finding the absolute minimal amount of data that needs to be collected to allow a proper statistical analysis of the central hypothesis. While clearly advisable to err on the side of caution (that is, collecting as many observations as possible), the price of excessive caution is a long data collection period with direct economic costs. Requiring that a single additional measurement be taken per finding, when multiplied by the thousands of findings in the experiment, adds up to a significant number of human hours. It is absolutely vital that a careful analysis of the statistical requirements be undertaken at the outset of a project.

The manner by which this experiment was conducted allowed for a midstream evaluation of the sufficiency of the data collected. After the Stanford component of the clinical experiment was completed, computer programs were developed to analyze the preliminary data. More will be said of these computer programs in the next section. They are mentioned here only because in the course of their development it was discovered that a slight modification of the experimental technique would greatly simplify the analysis. The simplification involved the selection of a different coordinate basis for measuring lesions. These modifications were made to the experimental design before beginning the University of Virginia component of the clinical experiment. The Stanford data were retroactively measured by students with respect to this new basis. This retroactive measuring process did not require any interaction with radiologists. It is solely responsible for the delays in completing the full lesion detection analysis on schedule. These analyses will be completed and submitted as an addendum to this report.

Data Analysis

The observations of the radiologists were captured by student assistants according to the instructions and on the forms listed in the previous section. Before the data could be analyzed it had to be entered into an electronic format suitable for processing by statistical programs developed for this experiment.

The World Wide Web was used for the process of entering the data. This medium offered several advantages: it allowed people to enter data concurrently and it allowed data to be entered from any computer with an Internet connection and a Web browser. Use was made of *forms*, a protocol for entering data through the Web. These forms proved to be a powerful medium for data entry. It was possible to check the data being entered to ensure it's logical consistency. For example, if a field required floating point numbers and something other than that was entered, a message would be shown to the user asking him or her to check the accuracy of their typing. The issue of accurate manual entry of data is extremely important. If the electronic version of the data doesn't match the version recorded during the clinical experiment, a garbage in/garbage out paradigm results. The Web in general and forms in particular tremendously reduced the likelihood of incorrect entry. It's use as a data entry facility is highly recommended.

That said, it is nearly impossible to ensure one hundred percent accuracy of manually entered data. Some measure of double data entry needs to be undertaken to check accuracy. This was indeed done in this experiment. Another method of data entry, designed both to reduce human labor and to increase accurate entry, is to make use of electronic scanners. There was no framework for using them in this study, i.e., no preexisting software base nor capital to acquire or develop such software. Retrospectively, scanning the paper forms instead of manually entering them would have been beneficial.

The programming language *perl* was heavily used in the course of this experiment. It provided the engine both for the entry of data on the Web (the form data is *posted* to a perl program) and for subsequent statistical analysis. All of the data collected in this project amounted to 2.38

megabytes of information. The data were stored in a moderately complex arrangement of flat text files. These text files were read and organized into a database suitable for searching and querying by a perl program. The perl database could be electronically searched; the information so generated was passed on to statistical programs (themselves often written in perl). Perl proved to be a cost effective solution to our database and analysis problems.

6.8 Relation to Original Statement of Work

The original Statement of Work is included as Appendix C. The original proposal requested three years of funding, but only two years were awarded and the proposal for completing and extending the original project was not successful. Hence the project was primarily devoted to the first two tasks.

As described previously, the data used in the study were acquired by digitizing analog mammograms as the digitally acquired mammograms available at the time were not deemed of sufficient quality. The primary Tasks 1 and 2 were carried out and completed for the comparisons of original analog, digital original, and lossy compressed digital mammograms using traditional film presentation. Task 3 was not carried out as it was originally intended for the third and unfunded year. Some progress was made towards algorithm development and work will continue on the theory and algorithm development with NSF support.

Data acquisition and clinical simulations required far more time than originally envisioned, largely because of the difficulties in conducting multi-institution and interdisciplinary work, the necessity for developing new statistical methods, and the logistical problems in the actual judging and gathering of experimental data. Nonetheless the clinical experiment comparing analog, digital, and lossy compressed digital mammograms was completed within the original budget by means of a no-cost extension.

7 Conclusions

The basic conclusions as detailed in the discussion in the previous sections are the following.

- A method of clinical experimental design and analysis for comparing analog, digital, and lossy compressed digital images has been developed, implemented, and tested in a pilot experiment involving 57 patients, 6 radiologist judges, and two additional radiologists serving as the expert panel for establishing independent gold standards. The method can be extended to other applications involving the comparison of images produced by differing image modalities or altered by computer processing. The pilot study was the largest data gathering experiment ever conducted to our knowledge for the comparison of analog and digital mammograms.

- The experiment demonstrated for the limited test set that for the resolution considered (50 micron spot size), lossy compression from 12 bits per pixel to .15 bits per pixel (80:1 compression ratio) results in no significant differences in management decisions based on the images among the analog, digital, or lossy compressed digital images.

- Preliminary experimental results demonstrated for the limited test set that for the resolution considered (50 micron spot size), lossy compression from 12 bits per pixel to .15 bits per pixel (80:1 compression ratio) results in no significant differences in lesion detection based on the images among the analog, digital, or lossy compressed digital images.

- Parameters for an experiment sufficiently large to be definitive in terms of size and power were estimated and reported.

- A new method for combining compression and classification for the purpose of automatically highlighting suspicious regions as part of the compression operation was extended to digital mammograms and the identification of microcalcifications and masses. The algorithm performs poorly in comparison to sophisticated pattern recognition techniques, but it is promising in that it involves no complicated signal processing, is incorporated into the compression algorithm, and is intended only to assist a radiologist by highlighting, not to make diagnoses. Given that the algorithm so far involves only operations on small 2×2 pixel blocks, it is expected that performance will improve considerably when the approach is extended to larger blocks and better probability distribution estimation methods.

8 Bibliography

References

[1] I. Andersson, "Mammography in clinical practice," *Med Radiography and Photography*, vol. 62, no. 2, p. 2, 1986.

[2] P. Stomper, J. Connolly, J. Meyer, and J. Harris, "Clinically occult ductal carcinoma in situ detected with mammography: analysis of 100 cases with radiographic-pathologic correlation," *Radiology*, vol. 172, pp. 235–241, 1989.

[3] D. Dershaw, A. Abramson, and D. Kinne, "Ductal carcinoma in situ: mammographic findings and clinical implications," *Radiology*, vol. 170, pp. 411–415, 1989.

[4] D. Ikeda, I. Andersson, C. W. rd L. Janzon, and F. Linell, "Radiographic appearance and prognostic consideration of interval carcinoma in the Malmö mammographic screening trial," *Amer. J. Roentgenology*, vol. 159, pp. 287–294, 1992.

[5] J. Martin, M. Moskowitz, and J. Milbrath, "Breast cancer missed by mammography," *Amer. J. Roentgenology*, vol. 132, pp. 737–739, 1979.

[6] D. Ikeda and I. Andersson, "Atypical mammographic presentations of ductal carcinoma in situ," *Radiology*, vol. 172, pp. 661–666, 1989.

[7] J. Frisell, G. Eklund, L. Hellstrom, and A. Somell, "Analysis of interval breast carcinomas in a randomized screening trial in Stockholm," *Breast Cancer Res Trea*, vol. 9, pp. 219–225, 1987.

[8] W. Kegelmeyer, "Software for image analysis aids in breast cancer detection," *OE Reports*, p. 7, February 1993.

[9] W. P. Kegelmeyer, Jr., "Evaluation of stellate lesion detection in a standard mammogram data set," in *Proceedings IS&T/SPIE Annual Symposium on Electronic Imaging, Science & Technology*, (San Jose, CA), Jan–Feb 1993.

[10] K. Woods, J. Solka, C. Priebe, C. Doss, K. Bowyer, and L. Clarke, "Comparative evaluation of pattern recognition techniques for detection of microcalcifications," in *Proceedings IS& T/SPIE Annual Symposium on Electronic Imaging, Science & Technology*, (San Jose, CA), January–February 1993.

[11] L. Clarke, G. Blaine, K. Doi, M. Yaffe, F. Shtern, and G. Brown, "Digital mammography, cancer screening: Factors important for image compression," in *Proceedings Space and Earth Science Data Compression Workshop*, (Snowbird, Utah), NASA, April 1993.

[12] W. B. Richardson, Jr., "Nonlinear filtering and multiscale texture discrimination for mammograms," in *Proceedings of SPIE*, vol. 1768 of *Mathematical Methods in Medical Imaging*, (San Diego, Calif.), pp. 293–305, SPIE, SPIE, July 1992.

[13] J. Sayre, D. R. Aberle, M. I. Boechat, T. R. Hall, H. K. Huang, B. K. Ho, P. Kashfian, and G. Rahbar, "Effect of data compression on diagnostic accuracy in digital hand and chest radiography," in *Proceedings of Medical Imaging VI: Image Capture, Formatting, and Display*, vol. 1653, pp. 232–240, SPIE, Feb. 1992.

[14] H. MacMahon, K. Doi, S. Sanada, S. Montner, M. Giger, C. Metz, N. Nakamori, F. Yin, X. Xu, H. Yonekawa, and H. Takeuchi, "Data compression: effect on diagnostic accuracy in digital chest radiographs," *Radiology*, vol. 178, pp. 175–179, 1991.

[15] H. Lee, A. H. Rowberg, M. S. Frank, H. S. Choi, and Y. Kim, "Subjective evaluation of compressed image quality," in *Proceedings of Medical Imaging VI: Image Capture, Formatting, and Display*, vol. 1653, pp. 241–251, SPIE, Feb. 1992.

[16] T. Ishigaki, S. Sakuma, M. Ikeda, Y. Itoh, M. Suzuki, and S. Iwai, "Clinical evaluation of irreversible image compression: Analysis of chest imaging with computed radiography," *Radiology*, vol. 175, pp. 739–743, 1990.

[17] K. Chan, S. Lou, and H. Huang, "Full-frame transform compression of CT and MR images," *Radiology*, vol. 171, no. 3, pp. 847–851, 1989.

[18] G. Wallace, "The JPEG still picture compression standard," *Communications of the ACM*, vol. 34, pp. 30–44, April 1991.

[19] M. Rabbani and P. W. Jones, *Digital Image Compression Techniques*, vol. TT7 of *Tutorial Texts in Optical Engineering*. Bellingham, WA: SPIE Optical Engineering Press, 1991.

[20] A. B. Watson, "Visually optimal DCT quantization matrices for individual mages," in *Proceedings of the 1993 IEEE Data Compression Conference (DCC)*, pp. 178–187, 1993.

[21] H. Blume, "ACR-NEMA Digital Imaging and Communication Standard Committee, Working Group # 4, MEDPACS Section, Data Compression Standard # PS2." 1989.

[22] J. W. Woods, ed., *Subband Image Coding*. Boston: Kluwer Academic Publishers, 1991.

[23] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," *IEEE Trans. Image Processing*, vol. 1, pp. 205–220, April 1992.

[24] T. Senoo and B. Girod, "Vector quantization for entropy coding of image subbands," *IEEE Transactions on Image Processing*, vol. 1, pp. 526–532, October 1992.

[25] A. S. Lewis and G. Knowles, "Image compression using the 2-D wavelet transform," *IEEE Trans. Image Processing*, vol. 1, pp. 244–250, April 1992.

[26] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[27] J. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3445–3462, December 1993.

[28] A. Said and W.A. Pearlman. A new fast and efficient image codec based on set partitioning in hierarchical trees. *IEEE Trans. on Circuits and Systems for Video Technology*, 1996. to appear.

[29] A. Said and W.A. Pearlman, "Set partitioning in hierarchical trees," `http://ipl.rpi.edu/SPIHT/`

[30] P. Wilhelm, D. R. Haynor, Y. Kim, and E. A. Riskin, "Lossy image compression for digital medical imaging system," *Optical Engineering*, vol. 30, pp. 1479–1485, Oct. 1991.

[31] H. H. Barrett, T. Gooley, K. Girodias, J. Rolland, T. White, and J. Yao, "Linear discriminants and image quality," in *Proceedings of the 1991 International Conference on Information Processing in Medical Imaging (IPMI '91)*, (Wye, United Kingdom), pp. 458–473, Springer-Verlag, July 1991.

[32] J. Bramble, L. Cook, M. Murphey, N. Martin, W. Anderson, and K. Hensley, "Image data compression in magnification hand radiographs," *Radiology*, vol. 170, pp. 133–136, 1989.

[33] M. Goldberg, M. Pivovarov, W. Mayo-Smith, M. Bhalla, J. Blickman, R. Bramson, G. Boland, H. Llewellyn, and E. Halpem, "Application of wavelet compression to digitized radiographs," *American Journal of Radiology*, vol. 163, pp. 463–468, August 1994.

[34] C. E. Metz, "Basic principles of ROC analysis," *Seminars in Nuclear Medicine*, vol. VIII, pp. 282–298, Oct. 1978.

[35] J. A. Swets, "ROC analysis applied to the evaluation of medical imaging techniques," *Investigative Radiology*, vol. 14, pp. 109–121, March–April 1979.

[36] J.A. Hanley and B.J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Diagnostic Radiology*, 143:29–36, 1982.

[37] B.J. McNeil and J.A. Hanley. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. *Medical Decision Making*, 4:137–150, 1984.

[38] S.J. Starr, C.E. Metz, L.B. Lusted, and D.J. Goodenough, "Visual detection and localization of radiographic images," *Radiology*, 1975; Vol. 116, pp. 533-538.

[39] D. Chakraborty and L. Winter, "Free-response methodology: alternate analysis and a new observer-performance experiment," *Radiology*, vol. 174, no. 3, pp. 873–881, 1990.

[40] H. C. Davidson, C. J. Bergin, C. Tseng, P. C. Cosman, L. E. Moses, R. A. Olshen, and R. M. Gray, "The effect of lossy compression on diagnostic accuracy of thoracic CT images." Presented at the 77th Scientific Assembly of the Radiological Society of North America, Chicago, Illinois, Dec. 1991.

[41] P. C. Cosman, H. C. Davidson, C. J. Bergin, C. Tseng,, L. E. Moses, R. A. Olshen, and R. M. Gray, "The effect of lossy compression on diagnostic accuracy of thoracic CT images," *Radiology*, vol. 190, no. 2, pp. 517–524, 1994.

[42] P. Cosman, C. Tseng, R. Gray, R. Olshen, L. E. Moses, H. C. Davidson, C. Bergin, and E. Riskin, "Tree-structured vector quantization of CT chest scans: image quality and diagnostic accuracy," *IEEE Trans. Medical Imaging*, vol. 12, pp. 727–739, Dec. 1993.

[43] S. Perlmutter, C. Tseng, P. Cosman, K. . Li, R. Olshen, and R. Gray, "Measurement accuracy as a measure of image quality in compressed MR chest scans," in *Proceedings of the IEEE 1994 International Symposium on Image Processing*, vol. 1, (San Antonio, Texas), pp. 861 – 865, October 1994.

[44] P. Cosman, R. Gray, and R. Olshen, "Evaluating quality of compressed medical images: SNR, subjective rating, and diagnostic accuracy," *Proceedings of the IEEE*, Volume 82, pp. 919–932, June 1994.

[45] R. M. Gray, R. A. Olshen, D. Ikeda, P. Cosman, S. Perlmutter, C. Nash, and K. Perlmutter, "Evaluating quality and utility in digital mammography," in *Proceedings of the 1995 IEEE International Conference on Image Processing*, IEEE, October 1995. volume II, pages 5–8, Washington, D.C., October 1995.

[46] R. M. Gray, R. A. Olshen, D. Ikeda, P.C. Cosman, S.M. Perlmutter, C. Nash, and K.O. Perlmutter "Measuring Quality in Computer Processed Radiological Images," *Proceedings of the Twenty Ninth Asilomar Conference on Signals, Systems, and Computers*, pp. 489–493, October 1995.

[47] C. N. Adams, A. Aiyer, B.J. Betts, J. Li, P. C. Cosman, S. M. Perlmutter, K. O. Perlmutter, D. Ikeda, L. Fajardo, R. Birdwell, B. L. Daniel, S. Rossiter, R. A. Olshen, and R. M. Gray, "Evaluating Quality and Utility of Digital Mammograms and Lossy Compressed Digital Mammograms," *Proceedings of the Third International Workshop on Digital Mammography*, Elsevier, Amsterdam, pages 169-176. (Preprint available at Web site [49].)

[48] S. M. Perlmutter, P. C. Cosman, R. M. Gray, R. A. Olshen, D. Ikeda, C. N. Adams, B.J. Betts, M. Williams, K. O. Perlmutter, J. Li, A. Aiyer, L. Fajardo, R. Birdwell, and B. L. Daniel "Image Quality in Lossy Compressed Digital Mammograms," *Signal Processing*, Special Issue on Medical Image Compression, to appear, 1997. (Preprint available at Web site [49].)

[49] Stanford University Compression and Classification Group, "USAMRMC Digital Mammography Image Quality Project," `http://www-isl.stanford.edu/~gray/army.html`, 1996.

[50] R.M. Gray, R.A. Olshen, D. Ikeda, P.C. Cosman, S.M. Perlmutter, C.L. Nash, and K.O. Perlmutter, "Full Breast Digital Imaging (FBDI) as a Screening Device: Summary of Clinical Protocol," Report submitted to Panel on Digital Mammography, FDA, 6 March 1995 and described in the transcription "Radiological Devices Panel Meeting" available from CASET Associates, Ltd., 10201 Lee Highway, Suite 160, Fairfax, VA 22030.

[51] W. Feller, *Introduction to Probability Theory*, third edition, Wiley, New York, 1968.

[52] P. Armitage, G. Berry, *Statistical Methods in Medical Research*, 3rd ed. Oxford, England: Blackwell, 1994.

[53] E. Lehmann, *Testing Statistical Hypotheses*, 2nd ed. New York: Wiley, 1986.

[54] A. Garber, R. Olshen, H. Zhang, and E. Venkatraman, "Predicting high-risk cholesterol levels," *International Statistical Review*, vol. 62, no. 2, pp. 203–228, 1994.

[55] G.W. Snedecor and W.G. Cochran, *Statistical Methods*, Iowa State University Press, Ames, Iowa, 1989.

[56] J. Cohen, "Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit," *Psychological Bulletin*, vol. 20, pp. 213–220, 1968.

[57] Y. Bishop, S. Feinberg, and P. Holland, *Discrete Multivariate Analysis.* Cambridge, Mass: MIT Press, 1975.

[58] J. Villasenor, B. Belzer, and J. Liao, "Wavelet filter evaluation for image compression," *IEEE Transactions on Image Processing*, Vol. 4, No. 8, August 1995, pp. 1053–60.

[59] I. H. Witten, R. M. Neal, and J. G. Cleary, "Arithmetic coding for data compression," *Communications of the ACM*, vol. 30, pp. 520–540, 1987.

[60] Vector quantization with zerotree significance map for wavelet image coding S. M. Perlmutter, K. O. Perlmutter, and P. C. Cosman *Proceedings of the Twenty Ninth Asilomar Conference on Signals, Systems, and Computers*, pp. 1419–1423, October 1995.

[61] P.C. Cosman, R.M. Gray, and M. Vetterli, "Vector quantization of image subbands: A survey," *IEEE Transactions on Image Processing*, vol.5, no.2, pp. 202-25. February, 1996.

[62] Jia Li, Navin Chaddha and Robert M. Gray, Multiresolution Tree Structured Vector Quantization, *Proc. Asilomar Conference on Signals, Systems and Computers*, Asilomar, California, Nov 1996.

[63] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth, Belmont, CA, 1984.

[64] K. L. Oehler, *Image Compression and Classification using Vector Quantization*, Ph.D. Dissertation, Stanford University Electrical Engineering Department, September 1993.

[65] K. L. Oehler and R. M. Gray, "Combining image classification and image compression using vector quantization", in *Proceedings of the 1993 IEEE Data Compression Conference (DCC)*, J.A. Storer and M. Cohn, Eds., Snowbird, Utah, March 1993, pp. 2–11, IEEE Computer Society Press.

[66] R. M. Gray, K. L. Oehler, K. O. Perlmutter, and R. A. Olshen, "Combining tree-structured vector quantization with classification and regression trees", in *Proceedings of the 27th Asilomar Conference on Circuits Systems and computers*, Pacific Grove, CA, 1993.

[67] K. O. Perlmutter, C. L. Nash, and R. M. Gray, "A comparison of bayes risk weighted vector quantization with posterior estimation with other VQ-based classifiers", in *IEEE International Conference on Image Processing*, Austin, TX, Nov. 1994, vol. 2, pp. 217–221.

[68] K. O. Perlmutter, *Compression and Classification of Images using Vector Quantization and Decision Trees*, Ph.D. Dissertation, Stanford University Electrical Engineering Department, December 1995.

[69] K. O. Perlmutter, R. M. Gray, R. A. Olshen, and S. M. Perlmutter, "Bayes risk weighted vector quantization with CART estimated class posteriors", in *PICASSP*, Detroit, MI, May 1995.

[70] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Press, 1992.

[71] T. Kohonen, J. Kangas, J. Laaksonen, and K. Torkkola, "LVQ_PAK: The learning vector quantization program package, version 2.1", Tech. Rep., Helsinki University of Technology, Laboratory of Computer and Information Science, Finland, Oct 1992, Available via anonymous ftp to cochlea.hut.fi (130.233.168.48).

A Illustrations, Figures, and Tables

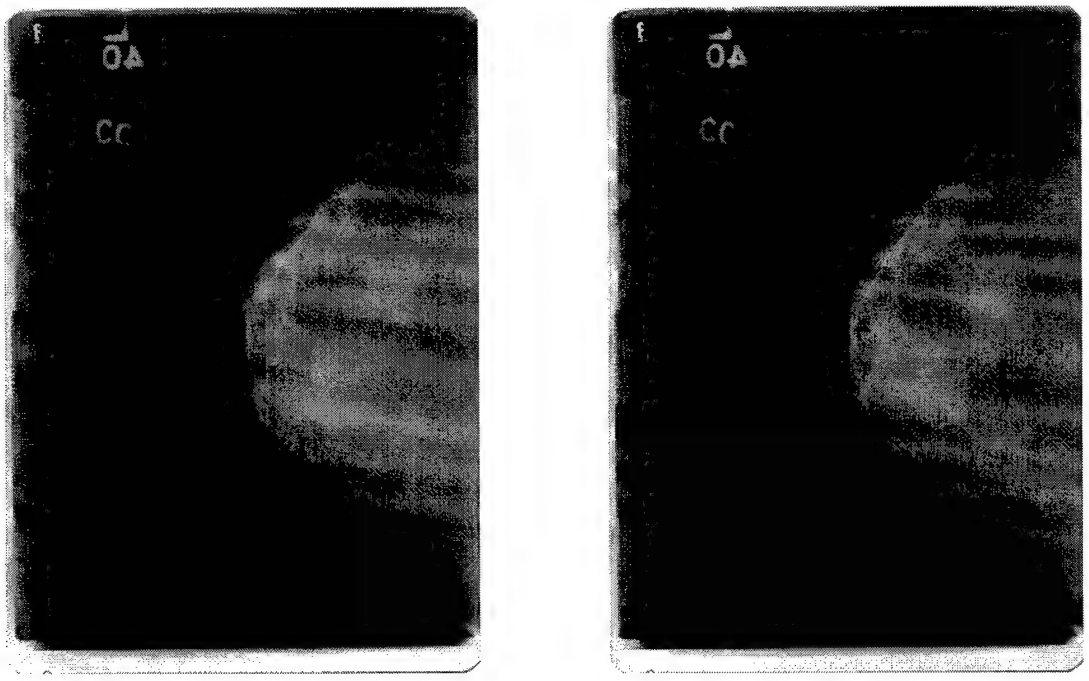


Figure 1: Original image and compressed image at 0.15 bpp in the ROI.

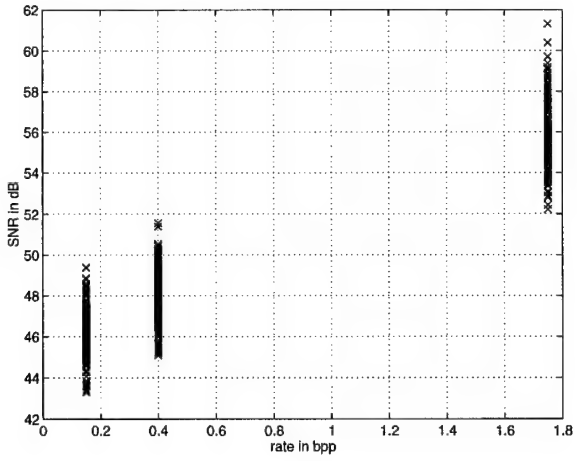


Figure 2: Scatter plot of SNR.

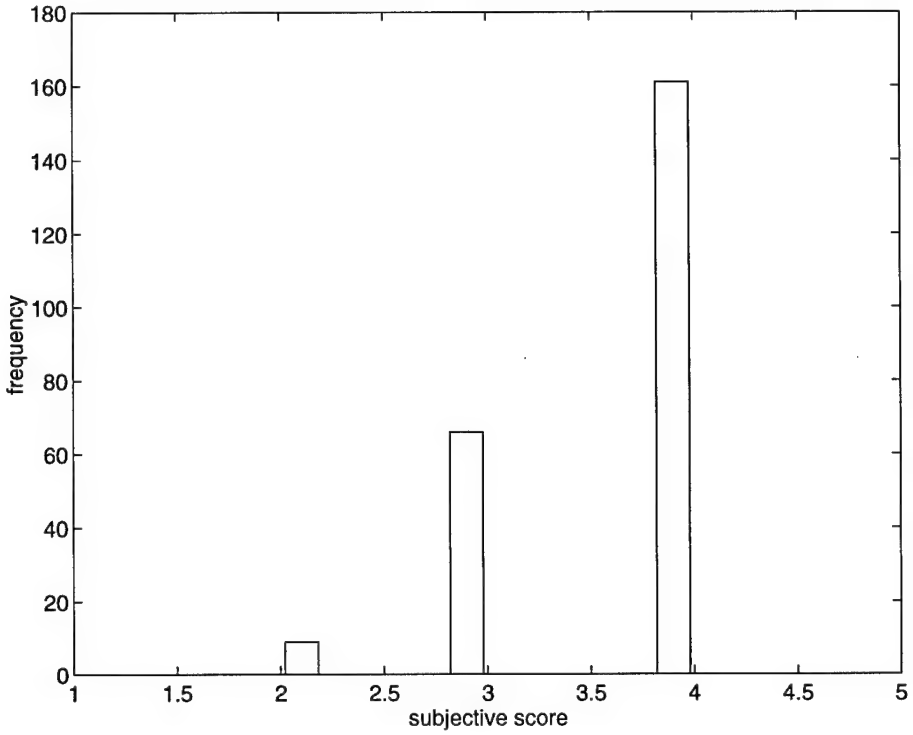


Figure 3: Subjective scores: analog images.

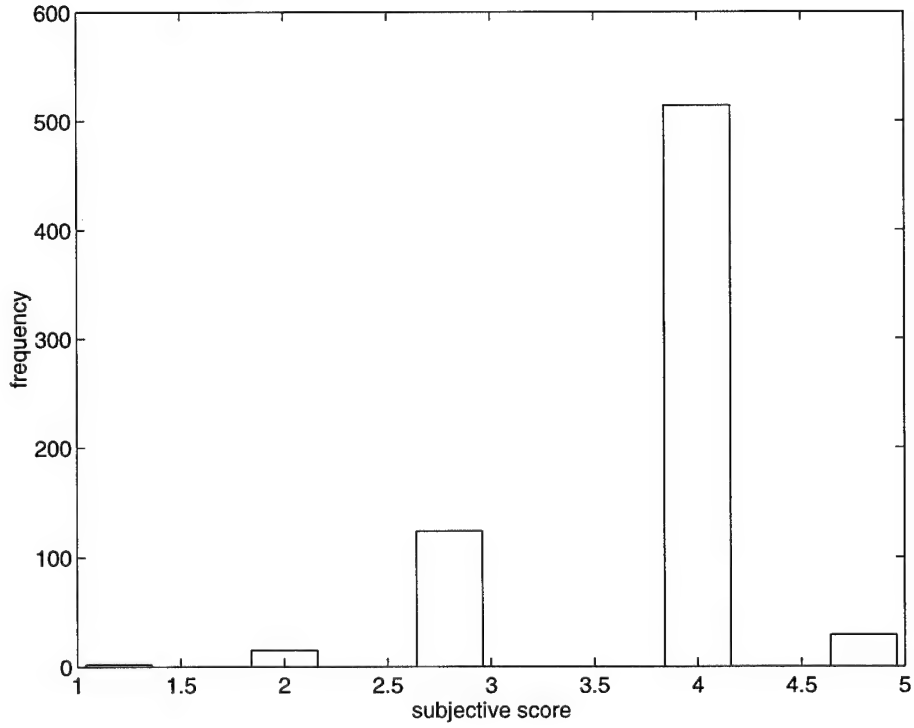


Figure 4: Subjective scores: original digital images.

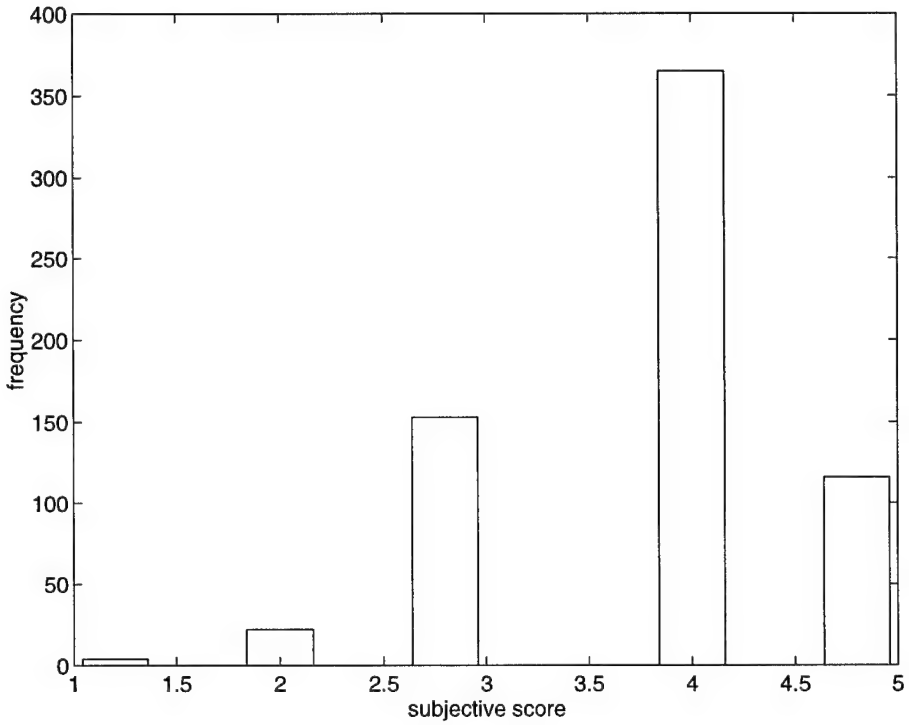
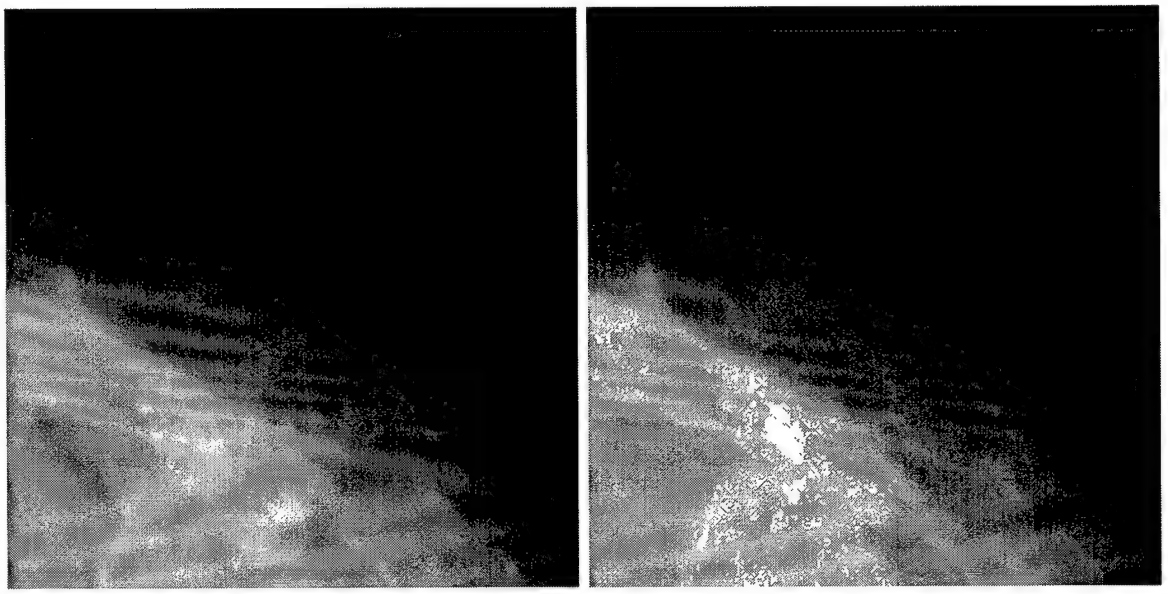
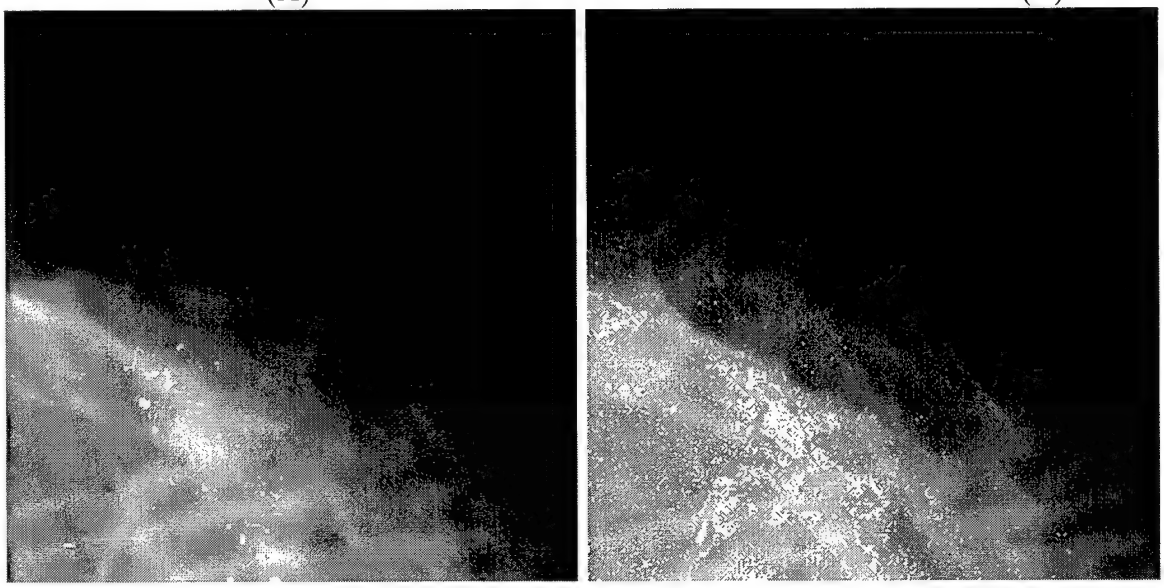


Figure 5: Subjective scores: lossy compressed digital images at 0.15 bpp.



(A) (B)



(C) (D)

Figure 6: Compression and classification of digitized mammograms at 2 bpp for calcifications: (A) Portion of Compressed Mammogram using BTSVQ with posterior estimation (B) Compressed/Classified image using BTSVQ with posterior estimation (white highlighted areas denote pixel blocks classified as microcalcifications) (C) Original 12 bit image with microcalcifications highlighted in white (D) Compressed/Classified image using independent TSVQ design (white highlighted pixel areas denote pixel blocks classified as microcalcification).

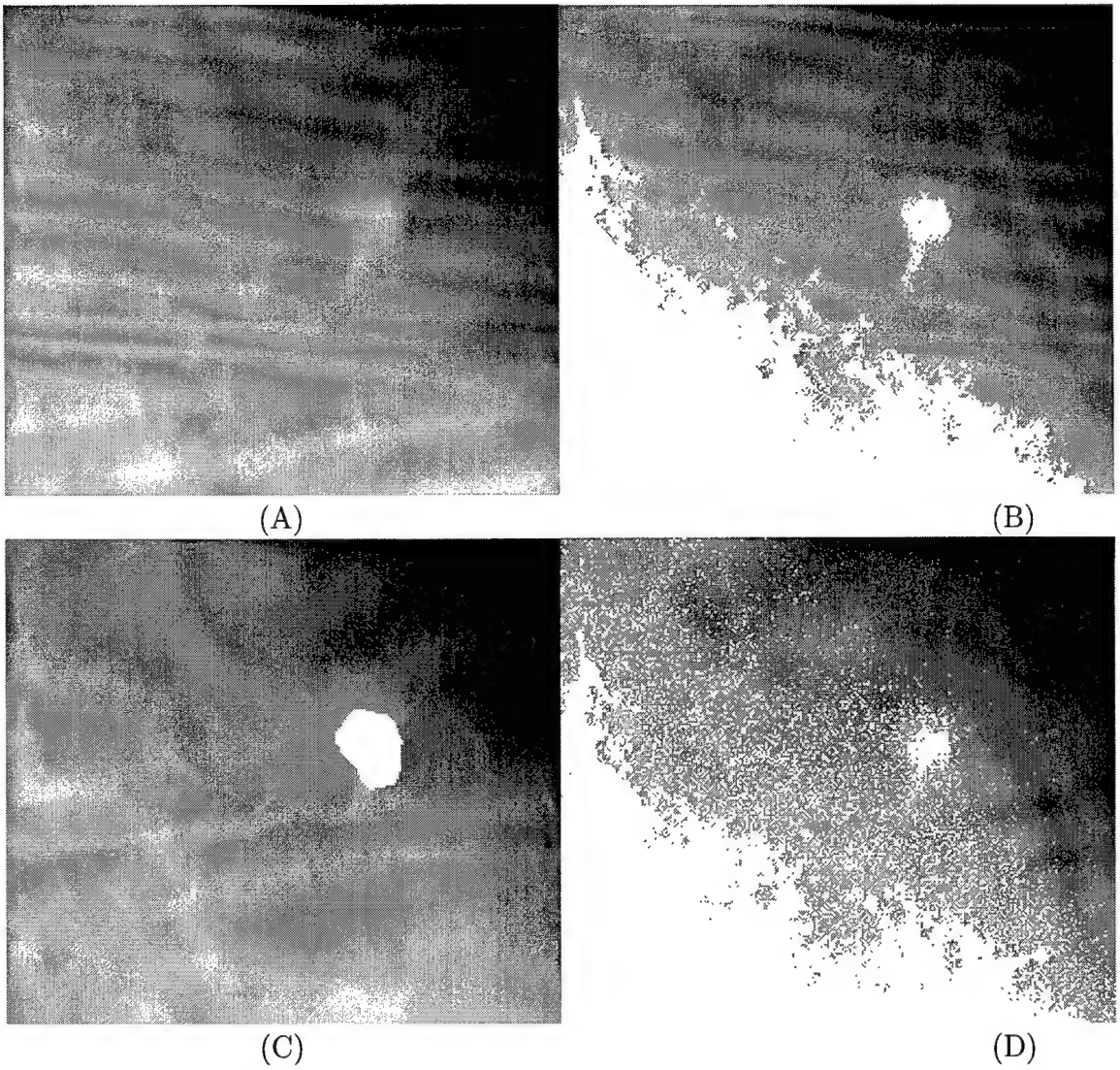


Figure 7: Compression and classification of digitized mammograms at 2 bpp for masses: (A) Portion of Compressed Mammogram (B) White area denotes highlighted pixel blocks classified as mass in (A) using posterior estimation (sensitivity 0.494, specificity 0.708) (C) White area denotes the actual mass area (D) White area denotes highlighted pixel blocks classified as mass in (A) using independent design (sensitivity 0.532, specificity 0.657).

6	benign mass
6	benign calcifications
5	malignant mass
6	malignant calcifications
3	malignant combination of mass & calcifications
3	benign combination of mass & calcifications
4	breast edema
4	malignant architectural distortion
2	malignant focal asymmetry
3	benign asymmetric density
15	normals

Table 1: Data Test Set: 57 studies, 4 views per study. Films printed using a Kodak 2180 X-ray film printer, a 79 micron 12 bit greyscale printer.

Category	Minimum number of cases
Total	400
Normal	200
Mammographically-detected breast cancers	110
Benign Findings	75
Breast Edemas	15

Table 2: Proposed data set.

II\ I	Right	Wrong	Row Sums
Right	$N(1,1)$	$N(1,2)$	$N(1,1)+N(1,2)$
Wrong	$N(2,1)$	$N(2,2)$	$N(2,1)+N(2,2)$
Column Sums	$N(1,1)+N(2,1)$	$N(1,2)+N(2,2)$	N

Table 3: Agreement 2×2 table.

II\ I	Right	Wrong	
Right	$2\psi + h - 1 + \gamma$	$1 - \psi - h - \gamma$	ψ
Wrong	$1 - \psi - \gamma$	γ	$1 - \psi$
	$\psi + h$	$1 - \psi - h$	1

Table 4: Management outcome probabilities.

II\ I	Right	Wrong	
Right	$a = \frac{N(1,1)}{N}$	$b = \frac{N(1,2)}{N}$	a+b
Wrong	$c = \frac{N(2,1)}{N}$	$d = \frac{N(2,2)}{N}$	c+b
	a+c	b+d	

Table 5: Agreement relative frequencies.

View	SNR		
	0.15 bpp ROI	0.4 bpp ROI	1.75 bpp ROI
left CC	45.93 dB	47.55 dB	55.30 dB
right CC	45.93 dB	47.47 dB	55.40 dB
left MLO	46.65 dB	48.49 dB	56.53 dB
right MLO	46.61 dB	48.35 dB	56.46 dB
left side (MLO and CC)	46.29 dB	48.02 dB	55.92 dB
right side (MLO and CC)	46.27 dB	47.91 dB	55.93 dB
Overall	46.28 dB	47.97 dB	55.92 dB

Table 6: Average SNR.

View	SNR, Bit Rate		
	0.15 bpp ROI	0.4 bpp ROI	1.75 bpp ROI
left CC	44.30 dB, 0.11 bpp	45.03 dB, 0.24 bpp	46.44 dB, 0.91 bpp
right CC	44.53 dB, 0.11 bpp	45.21 dB, 0.22 bpp	46.88 dB, 0.85 bpp
left MLO	44.91 dB, 0.11 bpp	45.73 dB, 0.25 bpp	47.28 dB, 1.00 bpp
right MLO	45.22 dB, 0.11 bpp	46.06 dB, 0.25 bpp	47.96 dB, 0.96 bpp
left side (MLO and CC)	44.60 dB, 0.11 bpp	45.38 dB, 0.24 bpp	46.89 dB ,0.96 bpp
right side (MLO and CC)	44.88 dB, 0.11 bpp	45.63 dB, 0.24 bpp	47.41 dB, 0.92 bpp
Overall	44.74 dB, 0.11 bpp	45.51 dB, 0.24 bpp	47.14 dB, 0.93 bpp

Table 7: Average SNR: full image, wavelet coding.

View	SNR		
	1.55 bpp ROI	1.90 bpp ROI	2.80 bpp ROI
left CC	42.51 dB	44.65 dB	50.68 dB
right CC	43.49 dB	45.78 dB	51.88 dB
left MLO	45.27 dB	47.38 dB	52.54 dB
right MLO	44.45 dB	46.73 dB	52.66 dB
left side (MLO and CC)	43.89 dB	46.02 dB	51.61 dB
right side (MLO and CC)	43.89 dB	46.25 dB	52.27 dB
Overall	43.93 dB	46.13 dB	51.94 dB

Table 8: Average SNR: ROI, perceptually optimized JPEG coding.

II\ I	R	W
R	7	2
W	1	2

RTS

II\ I	R	W
R	0	0
W	0	1

F/U

II\ I	R	W
R	6	4
W	3	5

C/B

II\ I	R	W
R	14	2
W	2	8

BX

(A) Analog vs. Digital Original

II\ I	R	W
R	6	3
W	1	2

RTS

II\ I	R	W
R	0	0
W	0	1

F/U

II\ I	R	W
R	8	2
W	2	6

C/B

II\ I	R	W
R	14	2
W	1	9

BX

(B) Analog vs. Digital Lossy Compressed: 1.75 bpp

II\ I	R	W
R	6	3
W	0	3

RTS

II\ I	R	W
R	0	0
W	0	1

F/U

II\ I	R	W
R	6	4
W	2	6

C/B

II\ I	R	W
R	12	3
W	4	6

BX

(C) Analog vs. Digital Lossy Compressed: 0.4 bpp

II\ I	R	W
R	4	4
W	0	3

RTS

II\ I	R	W
R	0	0
W	0	1

F/U

II\ I	R	W
R	3	7
W	4	4

C/B

II\ I	R	W
R	11	4
W	4	6

BX

(D) Analog vs. Digital Lossy Compressed: 0.15 bpp

Table 9: Agreement 2×2 tables for fadiologist A.

	RTS	F/U	C/B	BX
RTS	11	0	5	1
F/U	0	0	0	0
C/B	3	0	11	7
BX	2	0	2	15

	RTS	F/U	C/B	BX
RTS	4	0	0	0
F/U	0	0	0	1
C/B	3	0	3	3
BX	1	0	7	35

	RTS	F/U	C/B	BX
RTS	8	0	6	1
F/U	0	0	0	0
C/B	1	0	10	1
BX	0	0	7	23

A: Analog versus Digital

	RTS	F/U	C/B	BX
RTS	11	0	6	0
F/U	0	0	0	0
C/B	2	0	15	4
BX	1	0	2	16

	RTS	F/U	C/B	BX
RTS	2	1	0	1
F/U	0	1	0	0
C/B	3	1	3	2
BX	1	0	4	37

	RTS	F/U	C/B	BX
RTS	11	0	4	0
F/U	0	0	0	0
C/B	1	1	8	2
BX	1	0	5	24

B: Analog versus Lossy Compressed Digital: 1.75 bpp

	RTS	F/U	C/B	BX
RTS	9	0	6	2
F/U	0	0	0	0
C/B	1	0	10	10
BX	1	0	2	15

	RTS	F/U	C/B	BX
RTS	1	0	2	1
F/U	0	0	0	1
C/B	2	0	2	5
BX	2	0	5	36

	RTS	F/U	C/B	BX
RTS	7	0	7	1
F/U	0	0	0	0
C/B	2	0	8	2
BX	1	0	4	25

C: Analog versus Lossy Compressed Digital: 0.4 bpp

	RTS	F/U	C/B	BX
RTS	8	0	7	1
F/U	0	0	0	0
C/B	3	1	9	8
BX	1	0	6	11

	RTS	F/U	C/B	BX
RTS	3	1	0	0
F/U	0	1	0	0
C/B	3	0	3	2
BX	1	1	5	35

	RTS	F/U	C/B	BX
RTS	7	0	7	0
F/U	0	0	0	0
C/B	0	0	9	3
BX	0	0	9	20

D: Analog versus Lossy Compressed Digital: 0.15 bpp

Radiologist A Radiologist B Radiologist C

Table 10: Radiologist agreement tables, Stanford judges.

	RTS	F/U	C/B	BX
RTS	16	0	0	8
F/U	1	2	1	0
C/B	1	2	0	0
BX	2	1	1	22

	RTS	F/U	C/B	BX
RTS	22	0	5	0
F/U	0	0	3	0
C/B	5	1	7	1
BX	1	0	3	9

	RTS	F/U	C/B	BX
RTS	23	0	1	1
F/U	0	0	0	0
C/B	2	1	8	4
BX	0	0	3	14

A: Analog versus Digital

	RTS	F/U	C/B	BX
RTS	16	0	3	5
F/U	1	2	1	0
C/B	1	0	1	1
BX	0	0	2	24

	RTS	F/U	C/B	BX
RTS	24	0	2	1
F/U	0	0	3	0
C/B	5	0	7	2
BX	0	0	4	9

	RTS	F/U	C/B	BX
RTS	20	1	3	1
F/U	0	0	0	0
C/B	2	1	7	5
BX	1	0	0	16

B: Analog versus Lossy Compressed Digital: 1.75 bpp

	RTS	F/U	C/B	BX
RTS	15	1	2	6
F/U	3	1	0	0
C/B	1	1	1	0
BX	1	2	0	23

	RTS	F/U	C/B	BX
RTS	20	0	4	3
F/U	1	0	2	0
C/B	3	1	7	3
BX	0	1	3	9

	RTS	F/U	C/B	BX
RTS	20	1	3	1
F/U	0	0	0	0
C/B	2	0	7	6
BX	0	0	1	16

C: Analog versus Lossy Compressed Digital: 0.4 bpp

	RTS	F/U	C/B	BX
RTS	13	2	4	4
F/U	2	1	0	1
C/B	1	1	1	0
BX	1	0	1	24

	RTS	F/U	C/B	BX
RTS	19	1	4	2
F/U	0	0	2	1
C/B	6	0	7	1
BX	0	0	3	13

	RTS	F/U	C/B	BX
RTS	20	0	4	0
F/U	0	0	0	0
C/B	1	0	8	6
BX	0	0	2	15

D: Analog versus Lossy Compressed Digital: 0.15 bpp

Radiologist D Radiologist E Radiologist F

Table 11: Radiologist agreement tables, UVa judges.

level	judge	sensitivity		specificity		PVP
		mean	stdev	mean	stdev	mean
1	A	0.826	0.379	0.692	0.462	0.905
1	B	1.000	0.000	0.308	0.462	0.836
1	C	0.913	0.282	0.846	0.361	0.955
2	A	0.886	0.317	0.769	0.421	0.929
2	B	0.955	0.208	0.385	0.487	0.840
2	C	0.932	0.252	0.462	0.499	0.854
3	A	0.814	0.389	0.333	0.471	0.814
3	B	0.953	0.211	0.417	0.493	0.854
3	C	0.977	0.151	0.500	0.500	0.875
4	A	0.860	0.347	0.615	0.487	0.881
4	B	0.955	0.208	0.154	0.361	0.792
4	C	0.977	0.149	0.615	0.487	0.896
5	A	0.841	0.366	0.538	0.499	0.860
5	B	0.953	0.211	0.231	0.421	0.804
5	C	0.932	0.252	0.769	0.421	0.932

Table 12: Sensitivity, specificity and PVP for management: Level 1 refers to the analog images, level 2 to the uncompressed digital, and levels 3, 4, and 5 refer to those images where the breast section was compressed to 0.15, 0.4 and 1.75 bpp respectively (and where the label was compressed to .07 bpp). Only the Stanford judges are shown.

Table 13: Power as a function of parameters.

Power as a Function of Parameters				
ψ	γ	h	R	Power for a 5% Test
0.55	0.03	0.03	9	0.14
0.55	0.03	0.03	12	0.14
0.55	0.03	0.03	15	0.16
0.55	0.03	0.03	18	0.17
0.55	0.03	0.05	9	0.33
0.55	0.03	0.05	12	0.38
0.55	0.03	0.05	15	0.43
0.55	0.03	0.05	18	0.48
0.55	0.03	0.10	9	0.96
0.55	0.03	0.10	12	0.99
0.55	0.03	0.10	15	1.00
0.55	0.03	0.10	18	1.00
0.55	0.05	0.03	9	0.13
0.55	0.05	0.03	12	0.16
0.55	0.05	0.03	15	0.17
0.55	0.05	0.03	18	0.18
0.55	0.05	0.05	9	0.32
0.55	0.05	0.05	12	0.40
0.55	0.05	0.05	15	0.45
0.55	0.05	0.05	18	0.53
0.55	0.05	0.10	9	0.95
0.55	0.05	0.10	12	0.99
0.55	0.05	0.10	15	1.00
0.55	0.05	0.10	18	1.00
0.55	0.10	0.03	9	0.16
0.55	0.10	0.03	12	0.15
0.55	0.10	0.03	15	0.17
0.55	0.10	0.03	18	0.20
0.55	0.10	0.05	9	0.41
0.55	0.10	0.05	12	0.47
0.55	0.10	0.05	15	0.53
0.55	0.10	0.05	18	0.56
0.55	0.10	0.10	9	0.99
0.55	0.10	0.10	12	1.00
0.55	0.10	0.10	15	1.00
0.55	0.10	0.10	18	1.00
0.60	0.03	0.03	9	0.12
0.60	0.03	0.03	12	0.16
0.60	0.03	0.03	15	0.18
0.60	0.03	0.03	18	0.18
0.60	0.03	0.05	9	0.38
0.60	0.03	0.05	12	0.44
0.60	0.03	0.05	15	0.50
0.60	0.03	0.05	18	0.56
0.60	0.03	0.10	9	0.98
0.60	0.03	0.10	12	0.99
0.60	0.03	0.10	15	1.00
0.60	0.03	0.10	18	1.00
0.60	0.05	0.03	9	0.15
0.60	0.05	0.03	12	0.16
0.60	0.05	0.03	15	0.20
0.60	0.05	0.03	18	0.19
0.60	0.05	0.05	9	0.38
0.60	0.05	0.05	12	0.45
0.60	0.05	0.05	15	0.52
0.60	0.05	0.05	18	0.58
0.60	0.05	0.10	9	0.99
0.60	0.05	0.10	12	1.00
0.60	0.05	0.10	15	1.00

Continued

ψ	γ	h	R	Power for a 5% Test
0.60	0.05	0.10	18	1.00
0.60	0.10	0.03	9	0.15
0.60	0.10	0.03	12	0.19
0.60	0.10	0.03	15	0.21
0.60	0.10	0.03	18	0.22
0.60	0.10	0.05	9	0.45
0.60	0.10	0.05	12	0.55
0.60	0.10	0.05	15	0.62
0.60	0.10	0.05	18	0.67
0.60	0.10	0.10	9	1.00
0.60	0.10	0.10	12	1.00
0.60	0.10	0.10	15	1.00
0.60	0.10	0.10	18	1.00
0.65	0.03	0.03	9	0.16
0.65	0.03	0.03	12	0.18
0.65	0.03	0.03	15	0.18
0.65	0.03	0.03	18	0.22
0.65	0.03	0.05	9	0.42
0.65	0.03	0.05	12	0.50
0.65	0.03	0.05	15	0.56
0.65	0.03	0.05	18	0.63
0.65	0.03	0.10	9	0.99
0.65	0.03	0.10	12	1.00
0.65	0.03	0.10	15	1.00
0.65	0.03	0.10	18	1.00
0.65	0.05	0.03	9	0.15
0.65	0.05	0.03	12	0.18
0.65	0.05	0.03	15	0.22
0.65	0.05	0.03	18	0.23
0.65	0.05	0.05	9	0.48
0.65	0.05	0.05	12	0.54
0.65	0.05	0.05	15	0.64
0.65	0.05	0.05	18	0.67
0.65	0.05	0.10	9	1.00
0.65	0.05	0.10	12	1.00
0.65	0.05	0.10	15	1.00
0.65	0.05	0.10	18	1.00
0.65	0.10	0.03	9	0.19
0.65	0.10	0.03	12	0.24
0.65	0.10	0.03	15	0.25
0.65	0.10	0.03	18	0.28
0.65	0.10	0.05	9	0.58
0.65	0.10	0.05	12	0.66
0.65	0.10	0.05	15	0.72
0.65	0.10	0.05	18	0.81
0.65	0.10	0.10	9	1.00
0.65	0.10	0.10	12	1.00
0.65	0.10	0.10	15	1.00
0.65	0.10	0.10	18	1.00
0.70	0.03	0.03	9	0.20
0.70	0.03	0.03	12	0.19
0.70	0.03	0.03	15	0.23
0.70	0.03	0.03	18	0.28
0.70	0.03	0.05	9	0.54
0.70	0.03	0.05	12	0.61
0.70	0.03	0.05	15	0.68
0.70	0.03	0.05	18	0.77
0.70	0.03	0.10	9	1.00
0.70	0.03	0.10	12	1.00
0.70	0.03	0.10	15	1.00
0.70	0.03	0.10	18	1.00

Continued

ψ	γ	h	R	Power for a 5% Test
0.70	0.05	0.03	9	0.19
0.70	0.05	0.03	12	0.22
0.70	0.05	0.03	15	0.25
0.70	0.05	0.03	18	0.29
0.70	0.05	0.05	9	0.56
0.70	0.05	0.05	12	0.66
0.70	0.05	0.05	15	0.75
0.70	0.05	0.05	18	0.78
0.70	0.05	0.10	9	1.00
0.70	0.05	0.10	12	1.00
0.70	0.05	0.10	15	1.00
0.70	0.05	0.10	18	1.00
0.70	0.10	0.03	9	0.23
0.70	0.10	0.03	12	0.29
0.70	0.10	0.03	15	0.31
0.70	0.10	0.03	18	0.35
0.70	0.10	0.05	9	0.70
0.70	0.10	0.05	12	0.80
0.70	0.10	0.05	15	0.85
0.70	0.10	0.05	18	0.91
0.70	0.10	0.10	9	1.00
0.70	0.10	0.10	12	1.00
0.70	0.10	0.10	15	1.00
0.70	0.10	0.10	18	1.00
0.75	0.03	0.03	9	0.20
0.75	0.03	0.03	12	0.29
0.75	0.03	0.03	15	0.28
0.75	0.03	0.03	18	0.32
0.75	0.03	0.05	9	0.62
0.75	0.03	0.05	12	0.72
0.75	0.03	0.05	15	0.80
0.75	0.03	0.05	18	0.86
0.75	0.03	0.10	9	1.00
0.75	0.03	0.10	12	1.00
0.75	0.03	0.10	15	1.00
0.75	0.03	0.10	18	1.00
0.75	0.05	0.03	9	0.25
0.75	0.05	0.03	12	0.29
0.75	0.05	0.03	15	0.33
0.75	0.05	0.03	18	0.37
0.75	0.05	0.05	9	0.70
0.75	0.05	0.05	12	0.79
0.75	0.05	0.05	15	0.85
0.75	0.05	0.05	18	0.90
0.75	0.05	0.10	9	1.00
0.75	0.05	0.10	12	1.00
0.75	0.05	0.10	15	1.00
0.75	0.05	0.10	18	1.00
0.75	0.10	0.03	9	0.34
0.75	0.10	0.03	12	0.38
0.75	0.10	0.03	15	0.45
0.75	0.10	0.03	18	0.51
0.75	0.10	0.05	9	0.85
0.75	0.10	0.05	12	0.94
0.75	0.10	0.05	15	0.97
0.75	0.10	0.05	18	0.98
0.75	0.10	0.10	9	1.00
0.75	0.10	0.10	12	1.00
0.75	0.10	0.10	15	1.00
0.75	0.10	0.10	18	1.00
0.80	0.03	0.03	9	0.29

Continued

ψ	γ	h	R	Power for a 5% Test
0.80	0.03	0.03	12	0.34
0.80	0.03	0.03	15	0.40
0.80	0.03	0.03	18	0.45
0.80	0.03	0.05	9	0.79
0.80	0.03	0.05	12	0.87
0.80	0.03	0.05	15	0.93
0.80	0.03	0.05	18	0.96
0.80	0.03	0.10	9	1.00
0.80	0.03	0.10	12	1.00
0.80	0.03	0.10	15	1.00
0.80	0.03	0.10	18	1.00
0.80	0.05	0.03	9	0.33
0.80	0.05	0.03	12	0.39
0.80	0.05	0.03	15	0.44
0.80	0.05	0.03	18	0.50
0.80	0.05	0.05	9	0.85
0.80	0.05	0.05	12	0.92
0.80	0.05	0.05	15	0.96
0.80	0.05	0.05	18	0.98
0.80	0.05	0.10	9	1.00
0.80	0.05	0.10	12	1.00
0.80	0.05	0.10	15	1.00
0.80	0.05	0.10	18	1.00
0.80	0.10	0.03	9	0.55
0.80	0.10	0.03	12	0.62
0.80	0.10	0.03	15	0.68
0.80	0.10	0.03	18	0.74
0.80	0.10	0.05	9	0.99
0.80	0.10	0.05	12	0.99
0.80	0.10	0.05	15	1.00
0.80	0.10	0.05	18	1.00
0.85	0.03	0.03	9	0.44
0.85	0.03	0.03	12	0.52
0.85	0.03	0.03	15	0.57
0.85	0.03	0.03	18	0.62
0.85	0.03	0.05	9	0.96
0.85	0.03	0.05	12	0.98
0.85	0.03	0.05	15	0.99
0.85	0.03	0.05	18	1.00
0.85	0.03	0.10	9	1.00
0.85	0.03	0.10	12	1.00
0.85	0.03	0.10	15	1.00
0.85	0.03	0.10	18	1.00
0.85	0.05	0.03	9	0.54
0.85	0.05	0.03	12	0.61
0.85	0.05	0.03	15	0.69
0.85	0.05	0.03	18	0.74
0.85	0.05	0.05	9	0.99
0.85	0.05	0.05	12	0.99
0.85	0.05	0.05	15	1.00
0.85	0.05	0.05	18	1.00
0.85	0.05	0.10	9	1.00
0.85	0.05	0.10	12	1.00
0.85	0.05	0.10	15	1.00
0.85	0.05	0.10	18	1.00
0.85	0.10	0.03	9	0.93
0.85	0.10	0.03	12	0.97
0.85	0.10	0.03	15	0.99
0.85	0.10	0.03	18	1.00
0.85	0.10	0.05	9	1.00
0.85	0.10	0.05	12	1.00

level	judge	sensitivity		specificity	
		mean	stdev	mean	stdev
2	D	0.689	0.437	0.667	0.443
2	E	0.825	0.372	0.842	0.356
2	F	0.791	0.364	0.789	0.361
3	D	0.741	0.417	0.680	0.432
3	E	0.824	0.360	0.795	0.379
3	F	0.830	0.344	0.775	0.370
4	D	0.763	0.401	0.735	0.411
4	E	0.798	0.374	0.754	0.413
4	F	0.776	0.389	0.732	0.399
5	D	0.775	0.391	0.759	0.400
5	E	0.801	0.375	0.795	0.379
5	F	0.826	0.356	0.781	0.384
2	pooled	0.768	0.394	0.766	0.393
3	pooled	0.799	0.375	0.750	0.395
4	pooled	0.779	0.386	0.740	0.405
5	pooled	0.801	0.373	0.778	0.386

Table 14: Sensitivity, specificity and PVP for lesion detection: Level 2 refers to the uncompressed digital, and levels 3, 4, and 5 refer to those images where the breast section was compressed to 0.15, 0.4 and 1.75 bpp respectively (and where the label was compressed to .07 bpp). These numbers are with respect to the level 1 personal gold standard (that is, the analog originals). Only the UVa judges are shown.

level	judge	mean	stdev
1	A	3.90	.97
1	B	4.52	.75
1	C	4.59	.79
2	A	3.91	.41
2	B	3.85	.53
2	C	3.67	.65
3	A	3.82	.39
3	B	4.27	.93
3	C	3.49	.64
4	A	3.91	.39
4	B	3.93	.55
4	C	3.82	.50
5	A	3.92	.42
5	B	3.66	.57
5	C	3.82	.55
judges pooled			
1	pooled	4.33	.89
2	pooled	3.81	.55
3	pooled	3.86	.76
4	pooled	3.88	.49
5	pooled	3.80	.57

Table 15: Subjective scores.

Codebook design method:	SNR(dB)	Sensitivity	Specificity	Bayes Risk
Bayes TSVQ	29.2	41.19	92.60	0.106
Independent TSVQ	29.2	47.92	89.45	0.134
Kohonen's LVQ	19.3	59.48	55.08	0.471

Table 16: Statistical results of algorithms on mammogram images containing calcifications coded at 2 bpp.

Codebook design method:	SNR(dB)	Sensitivity	Specificity	Bayes Risk
Bayes TSVQ	32.96	49.45	70.80	0.351
Independent TSVQ	32.93	53.24	65.66	0.397

Table 17: Statistical results of algorithms on mammogram images containing masses coded at 2 bpp.

B Questionnaires/Clinical Protocols

This Appendix contains the basic observer form used in the clinical experiments and the instructions for the assistants who record the radiologist's decisions on the form and prompt for the ACR categories. The most recent versions of these forms are publically available at the project Web cite [49], along with "prompt sheets" the radiologists use for checking on possible categories.

ID number _____ Session number _____ Case number _____

Reader initials _____
Mammograms were of (**Left** **Right** **Both**) breast(s).

.....

Subjective rating for diagnostic quality:

(bad) 1 – 5 (good):

Left CC	Left MLO	Right CC	Right MLO

If any rating is < 4 the problem is:
1) sharpness 2) contrast 3) position 4) breast compression
5) noise 6) artifact 7) penetration

Recommend repeat? **Yes** **No**

Breast Density: Left **1** **2** **3** **4** Right **1** **2** **3** **4**

1) almost entirely fat 2) scattered fibroglandular densities
3) heterogeneously dense 4) extremely dense

Findings: **Yes** **No**

Note: If there are NO findings, the assessment is: **(1) (N) negative - return to screening**

• Findings (detection): **Dominant** **Incidental, focal** **Incidental, diffuse**

Individual finding side: **Left** **Right** **Both/Bilateral** Finding # _____ of _____

Finding type: **(possible, definite)**

1) mass
2) clustered calcifications
3) mass containing calcifications
4) mass with surrounding calcs
5) spiculated mass
6) ill defined mass
7) architectural distortion
8) solitary dilated duct
9) asymmetric breast tissue
10) focal asymmetric density
11) breast edema

12) multiple scattered and occasionally clustered benign appearing calcs
13) occasional scattered benign appearing calcs
14) multiple benign appearing masses
15) skin lesion
16) milk of calcium
17) plasma cell mastitis/secretory calcs
18) oil cysts
19) lymph node
20) fibroadenoma
21) calcified fibroadenoma
22) vascular calcs
23) dermal/skin calcs
24) post biopsy scar
25) reduction mammoplasty
26) implants
27) benign mass

28) other

Location:

1) UOQ **5) 12:00** **9) outer/lateral** **13) whole breast** **17) both breasts/ bilateral**
2) UIQ **6) 3:00** **10) inner/medial** **14) central**
3) LOQ **7) 6:00** **11) upper/cranial** **15) axillary tail**
4) LIQ **8) 9:00** **12) lower/inferior** **16) retroareolar**

View(s) in which finding is seen: **CC** **MLO** **CC and MLO**

Associated findings include: (p= possible, d= definite)

1) breast edema	**(p , d)**	**8) architectural distortion**	**(p , d)**
2) skin retraction	**(p , d)**	**9) calcs associated**	**(p , d)**
3) nipple retraction	**(p , d)**	**with mass**	
4) skin thickening	**(p , d)**	**10) multiple similar masses**	**(p , d)**
5) lymphadenopathy	**(p , d)**	**11) dilated veins**	**(p , d)**
6) trabecular thickening	**(p , d)**	**12) asymmetric density**	**(p , d)**
7) scar	**(p , d)**	**13) none**	**(p , d)**

Assessment: **The finding is**

(A) indeterminate/incomplete, additional assessment needed
What? 1) spot mag 2) extra views 3) U/S 4) old films 5) mag
What is your *best guess* as to the finding's 1–5 assessment? _____ or are you uncertain if the finding exists? *Y*

(1) (N) negative – return to screening

(2) (B) benign (also negative but with benign findings) – return to screening

(3) (P) probably benign finding requiring 6-month followup

(4L) (S) suspicion of malignancy (low), biopsy

(4M) (S) suspicion of malignancy (moderate), biopsy

(4H) (S) suspicion of malignancy (high), biopsy

(5) radiographic malignancy, biopsy

Comments: _____

_____

Measurements:

CC View Size: _____ cm long axis by _____ cm short axis
Distance from center of finding to: nipple _____ cm
left edge _____ cm, top edge _____ cm

MLO View Size: _____ cm long axis by _____ cm short axis
Distance from center of finding to: nipple _____ cm
left edge _____ cm, top edge _____ cm

ASSISTANT INSTRUCTIONS

PRE-SESSION INSTRUCTIONS

The radiologists will inform the assistants of suitable sites for hanging both the large digital films and the small analog films for this study.

Prior to the session, an assistant will either hang both the films and the clear overlays, or just the clear overlays for the session. Both film and overlay will need to be hung if this is the first sitting of this session. However, if the films are being re-used for another radiologist, they should have remained hanging and only the overlays will need to be hung.

After the films and the overlays are hung, the assistant should use a wax pencil and label the overlays in the upper outer portion of the image where the patient label is. You will put the session number, case number and reader initials. The most important information is the case number since it indicates which patient and which rate is being read. The other information can be added later, but the case number should be on the overlays prior to the session.

If there is a good place in the reading room to leave the envelopes for the films, that is fine. Otherwise, return the envelopes to the office. The films will be placed back into these envelopes after all the radiologist sittings for this session have occurred.

SESSION INSTRUCTIONS

Session Preparation

Make sure you have one *completed* cover sheet per case and many main sheets for the findings. You may want to take extra cover sheets, just in case something happens. You should also have your sheet of questions and a prompt sheet for the radiologist to reference.

The top portion of the cover sheet should be completed for each case in the session. If different people are hanging the films and conducting the session, you should coordinate the ordering of the films so that you do not have to hunt for the correct cover sheet during the session.

Besides the forms, you will need a pen to fill out the forms, a stapler and a magnifying glass for the radiologist. He/she may come with one, but we should have one for them, just in case. (This has been a request here, but the radiologists at your location may not have this same request.)

Session Prompting

1. Pause and let the radiologist have time to look over the case. They will give you some indication that they are ready to begin with the questioning for this patient.

2. Using the cover sheet for the case being viewed, complete the portion below the dotted line. Remember this is prompted for **ONCE** per patient case during the session. The following questions will be asked:

   (a) What is the quality of the films, 1 through 5, where 1 is bad and 5 is good?

   (b) Is this true for all views?

   YES - Write rating in all four boxes and continue

   NO - To radiologist: Please tell me the rating for each view.

   (c) If any view is rated 1,2 or 3 then ask:
   Please identify the reason you have given a rating less than four as being either sharpness, contrast, position, compression, noise, artifact or penetration.
   Would you recommend a repeat?

   (d) What is the breast density?

Notice that we

- Do not prompt for subjective rating for each view. We will obtain this information with the second question. If they answer 'no,' we will then ask for the rating for each view.
- Do not ask for the density of the left and right breasts individually. Typically the density will be the same in both. If it is not, the radiologist will automatically give you two densities.
- Do not read the density types, each radiologist will have a list to reference. If they give you a type not listed, simply let them know that they need to choose one of the options listed on their sheet.

3. Let's see if we need to go to the main section of the data forms. We will get this information with the following question: Are there any dominant or incidental findings?

NO - Circle *NO* for findings and proceed to the next case. Only the cover sheet will be completed for this case.

YES - Circle *YES* for findings.

- If they mention <u>one</u> type, go to that line of questioning.
- If they either do not mention the type of finding or they say that both types exists, go to the following question:
  Are there any specific, dominant, important findings?

  YES - go to QD.

  NO - go to QI and replace the question
  "What is it?" with "What is the incidental finding?"

QD <u>Dominant Findings</u>
Start filling out a findings sheet; this is the two page main section of the data sheet. There will be ONE sheet for each finding.

QD1 If the radiologist has NOT already done so, ask them:
Please outline the finding, number it and mark the nipple.

- Many times the radiologist will outline the area while they are getting familiar with the image. However, they may not mark the nipple or number the finding(s). In addition, if there are multiple findings on one side, the breast marking will only be asked once.

(a) What and where is the finding?

(b) Is it definite or possible?

(c) What views do you see it in?

(d) Are there any (other) associated findings?

YES - What is it? Is it possible or definite?

- If the associated finding is "calcs associated with mass" say:
  Please label the finding as xA and the associated finding as xB (where x is the finding number). Please outline the associated finding.
- Repeat.

NO - Continue.

(e) What is your assessment?
indeterminate or incomplete - ask the following

- What would you request?

- What is your best guess of an assessment? <pause> or are you uncertain if the finding exists?

(f) Are there any other specific, dominant, important findings?

YES - Return to QD1 and begin the questions for the next finding.

NO - Go to QI.

QI Incidental Findings Start filling out a findings sheet; this is the two page main section of the data sheet. There will be ONE sheet for each finding.

QI1 Are there (other) incidental findings that you would mention in your report, either focal or diffuse?

NO - Staple forms to the cover sheet and proceed to the next case.

YES & FOCAL

(a) What and where is it?

(b) Is it definite or possible?

(c) If they have NOT already done so, ask the radiologist:
Please outline the incidental finding, number it and mark the nipple.

(d) What views do you see it in?

(e) Are there any (other) associated findings?

YES - What is it? Is it possible or definite?

- If the associated finding is "calcs associated with mass" say:
Please label the finding as xA and the associated finding as xB (where x is the finding number). Please outline the associated finding.
- Repeat

NO - Continue.

(f) Are there any comments accompanying the *benign* assessment you would include in your report?

(g) Return to QI1.

YES & DIFFUSE

(a) What and where is it?

(b) Is it definite or possible?
They should indicate the finding side, the type, location and views. If any of these are not answered, ask for specifics.

(c) Are there any comments accompanying the *benign* assessment you would include in your report?

(d) Return to QI1.

POST-SESSION INSTRUCTIONS

After the session, measure the size , the distance to the nipple and the horizontal and vertical offsets of the findings. The measurements will be recorded at the end of the findings sheet, the main section of the data form. The only findings that are not measured are the diffuse incidental findings.

Place the overlays and the data package in envelopes for safe keeping. There should be one envelope per case. Remember to bring a set of envelopes for the overlays.

If these films are to be viewed by another radiologist, leave them hanging. If this is the last session with these films, take them down and return them to their designated envelope. The envelopes for the films may have been left in a designated area. However, if they were not, remember to bring these envelopes with you.

C Original Statement of Work

Compression and Classification of Digital Mammograms for Storage, Transmission, and Computer Aided Screening

Task 1, Acquisition of Data and Compression Algorithm Selection, Months 1–12:

a. Expand our existing database of digitized analog mammograms. Obtain digitally acquired mammograms (spot mammograms in the near future and full frame as available).

b. Label database: marking of microcalcifications and masses on hardcopy, corresponding pixel-by-pixel labeling on computer. This has already been done for our current database.

c. Compare several compression algorithms in terms of easily found quality measures, signal-to-noise ratios vs. bit rate, informal opinions of radiologists

d. Finalize details of compression study: monitor dynamic range, windows and levels, zoom, image randomization, marking protocol.

e. Design of finite-state VQs and classifiers for important features based on empirical Bayes, CART, and heuristic methods. Determine good classifiers operating on small pixel blocks for use in both compression and combined compression/classification.

Task 2, Clinical simulations and statistical analysis for compression, Months 13–30:

a. Perform the clinical simulation based on the findings/assessment/management questionnaire with four judges for the selected compression methods. Obtain independent gold standard from 2 additional judges.

b. Tabulate data and perform planned statistical analyses to compare bit rates, judges, and film vs. monitor for both personal and independent gold standards.

c. Resolve any new issues arising from the data by additional statistical analyses as necessary.

d. Study alternative computable distortion measures: Use SNR and other computable measures proposed in the literature to predict the diagnostic accuracy (sensitivity, predictive value positive, measurement accuracy) and the subjective ratings data.

Task 3, Computer Assisted Diagnosis/Screening simulations and statistical analysis, Months 18–36:

a. Pilot study for combined compression/classification/enhancement: Simulate screening environment with high resolution monitors capable of false color display. Study informally the relative merits of various highlighting and enhancement algorithms in order to select most promising for full study.

b. Full clinical simulation based on the findings/assessment/management questionnaire with four judges for the selected compression/classification methods providing optional instantaneous highlighting of regions classified as abnormal. Possible incorporation of other signal processing.

c. Statistical analysis of clinical utility of classification relative to independent gold standard.

d. Statistical analysis of clinical utility of classification relative to biopsy results.

e. Incorporate biopsy information into labeling: Redesign codebooks using improved class posterior probabilities based on training data labeled using biopsy results as well as original radiologist labels. Automatic reclassification of test data using new codes and comparison of performance against previous codes.

D Publications supported by this grant.

1. C.L. Nash, K.O. Perlmutter, and R.M. Gray "Evaluation of Bayes risk weighted vector quantization with posterior estimation in the detection of lesions in digitized mammograms." *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers.* Held: Pacific Grove, CA, USA, 31 Oct.-2 Nov. 1994. (USA: IEEE Comput. Soc. Press, 1994. p. 716-20 vol.1)

2. R. M. Gray, R. A. Olshen, D. Ikeda, P. Cosman, S. Perlmutter, C. Nash, and K. Perlmutter, "Evaluating quality and utility in digital mammography," in *Proceedings of the 1995 IEEE International Conference on Image Processing*, IEEE ICIP '95, October 1995. volume II, pages 5–8, Washington, D.C., October 1995.

3. R. M. Gray, R. A. Olshen, D. Ikeda , P.C. Cosman, S.M. Perlmutter, C. Nash, and K.O. Perlmutter "Measuring Quality in Computer Processed Radiological Images," *Proceedings of the Twenty Ninth Asilomar Conference on Signals, Systems, and Computers*, pp. 489–493, October 1995.

4. Vector quantization with zerotree significance map for wavelet image coding S. M. Perlmutter, K. O. Perlmutter, and P. C. Cosman *Proceedings of the Twenty Ninth Asilomar Conference on Signals, Systems, and Computers*, pp. 1419–1423, October 1995.

5. K. O. Perlmutter, S. M. Perlmutter, R. M. Gray, R. A. Olshen, K. L. Oehler, "Bayes Risk Weighted Vector Quantization with Posterior Estimation for Image Compression and Classification," *IEEE Transactions on Image Processing*, vol.5, no.2, p. 347-60, February 1996.

6. P.C. Cosman, R.M. Gray, and M. Vetterli, "Vector quantization of image subbands: A survey," *IEEE Transactions on Image Processing*, vol.5, no.2, pp. 202-25. February, 1996.

7. C. N. Adams, A. Aiyer, B.J. Betts, J. Li, P. C. Cosman, S. M. Perlmutter, K. O. Perlmutter, D. Ikeda, L. Fajardo, R. Birdwell, B. L. Daniel, S. Rossiter, R. A. Olshen, and R. M. Gray, "Evaluating Quality and Utility of Digital Mammograms and Lossy Compressed Digital Mammograms," *Proceedings of the Third International Workshop on Digital Mammography*, Elsevier, Workshop held June 1996, Amsterdam, pages 169–176. (Preprint available at Web site [49].)

8. Jia Li, Navin Chaddha and Robert M. Gray, Multiresolution Tree Structured Vector Quantization, *Proc. Asilomar Conference on Signals, Systems and Computers*, Asilomar, California, Conference held Nov 1996, *Proceedings* to appear.

9. S. M. Perlmutter, P. C. Cosman, R. M. Gray, R. A. Olshen, D. Ikeda, C. N. Adams, B.J. Betts, M. Williams, K. O. Perlmutter, J. Li, A. Aiyer, L. Fajardo, R. Birdwell, and B. L. Daniel "Image Quality in Lossy Compressed Digital Mammograms," *Signal Processing*, Special Issue on Medical Image Compression, to appear, 1997. (Preprint available at Web site [49].)

Invited Plenary Lecture "Measuring quality and utility in lossy compressed medical images," Robert M. Gray, 1995 IEEE Data Compression Conference (DCC), Snowbird Utah, 28 March 1995. (Plenary talks are listed in the conference program but not in the *Proceedings* because they are invited, not refereed, papers.)

One or two summary papers will be written and submitted for publication during the spring and summer following the completion of several statistical analyses of the data that we are still working on. Several of the yet unpublished results reported in this report will be included. We have not chosen to limit the distribution of this report because of the inclusion of unpublished results. We have no objection to the results being distributed provided this report is cited as their origin.

E List of Personnel

The following people participated in the project during its life:

1. Robert M. Gray, PhD, Principal Investigator. Professor and Vice Chair of Electrical Engineering, Stanford University.
2. Richard Olshen, PhD, Investigator. Professor of Health Research and Policy, and, by courtesy, of Statistics and of Electrical Engineering, Stanford University.
3. Debra Ikeda, MD, Investigator. Former Assistant Professor and Chief, Breast Imaging Section, Department of Radiology, Stanford University.
4. Robin Birdwell, MD, Investigator. Assistant Professor and acting Chief, Breast Imaging Section, Department of Radiology, Stanford University.
5. Pamela C. Cosman, PhD, Former Research Assistant and Post Doctoral Fellow, Stanford University. Now Assistant Professor of Electrical Engineering, University of California at San Diego.
6. Sharon M. Perlmutter, PhD, Former Research Assistant, Stanford University. Now Research Engineering, Johnson-Grace Inc., Newport Beach, Calif.
7. Keren O. Perlmutter, PhD, Former Research Assistant, Stanford University. Now Research Engineering, Johnson-Grace Inc., Newport Beach, Calif.
8. Bradley J. Betts, MS, Research Assistant, Department of Electrical Engineering, Stanford University.
9. Cheryl Nash Adams, MS, Research Assistant, Department of Electrical Engineering, Stanford University.
10. Jia Li, MS, Research Assistant, Department of Electrical Engineering, Stanford University.
11. Anuradha Aiyer, MS, Research Assistant, Department of Electrical Engineering, Stanford University.
12. Sarah Horine, Undergraduate Student Technical Assistant, Stanford University.
13. Robin Baxter, Technical Assistant at the University of Virginia

In addition to the paid staff several people devoted volunteer time:

1. Bruce L. Daniel, MD. Fellow of Radiology, Stanford University.
2. Laurie L. Fajardo, MD. Professor and Vice Chair for Research, Department of Radiology, University of Virginia.
3. Mark Williams, PhD. Assistant Professor of Radiology, Physics, and Biomedical Engineering, University of Virginia.
4. Stanford Rossiter, M.D. Staff Physician, Breast Imaging Section, Department of Radiology, Stanford University.
5. Edward Sickles, M.D. Professor of Radiology, University of California at San Francisco, and Chief of Radiology, Mt. Zion Hospital.

6. R. Moran, M.D. Assistant Professor of Radiology, University of Virginia.
7. Gia DeAngelis M.D. Assistant Professor of Radiology, University of Virginia.
8. Dalia Gomez, Undergraduate Student Technical Assistant, Stanford University.